Spring 2010

Characterization of Protein Residue Surface Accessibility Using Sequence Homology

Radhika Pallavi Mishra
San Jose State University

CHARACTERIZATION OF PROTEIN RESIDUE SURFACE ACCESSIBILITY USING SEQUENCE HOMOLOGY

A Thesis

Presented to

The Faculty of the Department of Chemistry

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Radhika Pallavi Mishra

May 2010

The Designated Thesis Committee Approves the Thesis Titled

CHARACTERIZATION OF PROTEIN RESIDUE SURFACE ACCESSIBILITY
USING SEQUENCE HOMOLOGY

by

Radhika Pallavi Mishra

APPROVED FOR THE DEPARTMENT OF CHEMISTRY

SAN JOSE STATE UNIVERSITY

May 2010

Dr. Brooke Lustig	Department of Chemistry
Dr. Elaine D. Collins	Department of Chemistry
Dr. Marc d'Alarcao	Department of Chemistry

ABSTRACT

CHARACTERIZATION OF PROTEIN RESIDUE SURFACE ACCESSIBILITY USING SEQUENCE HOMOLOGY

by Radhika Pallavi Mishra

Residues present on the surface of the proteins are involved in a number of functions, especially in ligand-protein interactions, that are important for drug design. The residues present in the core of the protein provide stability to the protein and help in maintaining protein structure. Hence, there is a need for a binary characterization of protein residues based on their surface accessibility (surface accessible or buried). Such a classification can aid in the directed study of either residue type.

A number of methods for the prediction of surface accessible protein residues have been proposed in the past. However, most of these methods are computationally complex and time consuming. In this thesis, we propose a simple method based on protein sequence homology parameters for the binary classification of protein residues as surface accessible or “buried”. To aid in the classification of protein residues, we chose three highly conservative homology-based parameter filter thresholds. The filter thresholds predicted and evaluated are: residue sequence entropy ≥ 0.15, fraction of strongly hydrophobic residues < 0.5 and fraction of small residues < 0.15. The application of these filter thresholds to the residues, is expected to predict the “buried residues” with a better percentage accuracy than that of the surface accessible residues.

These filter thresholds were selected from the frequency distributions and the aggregate correlation plots of the various homology-based parameters. An analysis of the plots suggests the presence of a strongly hydrophobic core between packing density $14 - 22$ where the presence of strongly hydrophobic residues is maximum and the presence of small and non-strongly hydrophobic residues is minimum. However, the densest portion of the protein (density $26 - 35$) is indicated to be occupied by a combination of small and non-strongly hydrophobic residues with a negligible presence of strongly hydrophobic residues.

ACKNOWLEDGEMENT

First of all, I would like to express my gratitude towards my research advisor, Dr. Brooke Lustig, for patiently guiding me through this work. I would like to thank him for being kind and understanding in allowing me to be on a flexible work routine and always being available for lengthy discussions. I am deeply indebted to him for introducing me to bioinformatics and consistently guiding, encouraging and supporting me all through my work. I would also like to thank my M. S. thesis committee members, Dr. Elaine D. Collins and Dr. Marc d'Alarcao, for spending their valuable time on reviewing my thesis and providing insightful feedback. I would also like to thank my husband, Kesari Mishra, for his able support and my son, Sanatan, for his cooperations during my entire study. In the end, I would like to extend my heartfelt gratitude to my parents for their blessings and encouragement.

Contents

List of Figures **ix**

List of Tables **xi**

1 Introduction **1**

1.1 Existing Approaches 2

1.1.1 Experimental Methods 2

1.1.2 Protein Structure Modeling Approaches 3

1.1.2.1 Comparative Modeling 3

1.1.2.2 Threading 4

1.1.2.3 *Ab Initio or de Novo* 4

1.1.3 Sequence Homology Methods 5

1.2 Important Terms 6

1.2.1 Packing Density 6

1.2.2 Sequence Entropy 7

1.2.3 Surface Accessibility 8

1.3 Definitions 9

1.4 Overview 11

1.5 Organization of the Thesis 12

2 Methods **14**

2.1 Protein Set Preparation 14

2.2 Residue Packing Density 15

2.3 Sequence Variability 16

2.3.1 Calculation of Sequence Entropy 17

2.3.2 6-point Entropy 18

2.3.3 Calculation of HSSP Entropy 19

2.4 Fractional Calculations 20

2.5 PDB-FASTA Reconciliation 21

2.6 Aggregate Analysis 22

2.6.1 Correlation Plots 22

2.7 Frequency Distributions 23

2.8 Characterization of Protein Lists 23

3 Results 24

3.1 Characterization of Protein Lists 24

3.1.1 Characterization of Learning Set List 24

3.1.2 Characterization of Monomeric List 36

3.1.3 Characterization of Homodimeric List 41

3.1.4 Characterization of Heterodimeric List 47

3.2 Aggregate Correlation Plots 52

3.2.1 Validation of Packing Density Calculation 52

3.2.2 Validation of Entropy Calculations 58

3.2.3 Entropy Inverse-Density Correlation Plots 58

3.3 Frequency Distributions 72

3.3.1 Density Distribution 72

3.3.2 Bin Distribution 73

3.3.3 Filter Parameter Based Analysis 81

4 Discussion 87

4.1 Packing Density: A Measure of Compactness in a Protein 87

4.2 Aggregate Trend Analysis 88

4.2.1 Sequence Entropy: A Measure of Flexibility or Mutability . . . 88

4.2.2 Fractional Parameters 89

4.3 Filter Threshold Selection And Surface Accessibility Prediction 92

5 Conclusions 96

6 Future Studies 98

Bibliography 100

Appendices

A Program Listings 107

B Additional Figures 132

C Additional Tables 142

List of Figures

3.1 Frequency distributions for the characterization of learning set list of proteins . . . 35

3.2 Frequency distribution plots for the characterisation of monomeric list of proteins . . . 40

3.3 Frequency distribution plots for the characterization of homodimeric list of proteins . . . 46

3.4 Frequency distribution plots for the characterization of heterodimeric list of protein chains . . . 51

3.5 Choice of hydrophobicity scale and parameters . . . 54

3.6 Hydrophobic hydrophilic fraction comparison . . . 56

3.7 Density-relative surface accessibility comparison . . . 57

3.8 Various entropy comparison . . . 59

3.9 Aggregate entropy-inverse density trend comparison for the four protein lists . . . 61

3.10 Aggregate fraction strongly hydrophobic-inverse density trend comparison for the four protein lists . . . 64

3.11 Aggregate fraction small residues-inverse density trend comparison for the four protein lists . . . 66

3.12 Aggregate fraction non-strongly hydrophobic vs inverse packing density trend comparison for the four protein lists . . . 68

3.13 Aggregate fraction gaps versus inverse density trend comparison for the four protein lists . . . 70

3.14 Aggregate correlation plots of various homology-based parameters and inverse packing density . . . 71

3.15 Density distribution plots for the fractional parameters . . . 74

3.16 Overlay plot of density distributions for various fractional parameters . 75

3.17 Bin frequency distributions of various fractional parameters for the learning set list 79

3.18 Overlay frequency distributions of learning set list for various fractional parameters 80

3.19 Threshold frequency distributions for entropy 82

3.20 Threshold frequency distributions for Fraction Small Residues (FSR) . . 84

3.21 Threshold frequency distributions for Fraction Strongly Hydrophobic (FSHP) 86

B.1 Aggregate entropy-inverse density correlation plots of 268 proteins for a set of BLASTP alignment cutoffs. 132

B.2 Entropy-inverse density correlation plots. 133

B.3 Gaps Included Aggregate plots for learning set list. 134

B.4 Entropy inverse density plots for a set of 268 proteins. 135

B.5 Z Score analysis of the learning set list of proteins. 136

B.6 Aggregate correlation plot of fraction strongly hydrophobic for the learning set list. 137

B.7 Visual alignment of 1LS9 and 1C6S. 138

B.8 Alignment of 1C6S and 1LS9, two related proteins by ClustalW in Jalview 2.4. 139

B.9 RSA versus Fraction strongly hydrophobic for the learning set list of 268 proteins. 140

B.10 Frequency distribution of query residues for the learning set list of 268 proteins. 141

List of Tables

3.1 Learning set list of 268 proteins 25

3.2 Query length and number of alignments of individual proteins for the learning set list of 268 proteins. 32

3.3 Monomeric List of 75 proteins 36

3.4 Query length and number of alignments of individual proteins for the list of 75 monomeric proteins. 39

3.5 Homodimeric list of 106 proteins 42

3.6 Query length and number of alignments of individual proteins for the list of 106 homodimeric proteins 45

3.7 Heterodimeric list of 50 protein chains 47

3.8 Query length and number of alignments of individual proteins for the list of 50 heterodimeric protein chains. 49

C.1 List of 65 monomeric protein chains present in the learning set list. . . 142

C.2 List of 99 homodimeric protein chains present in the learning set list. . 142

C.3 List of 40 heterodimeric protein chains present in the learning set list. . 142

C.4 Z Score values at each density position for the learning set list of proteins. 143

C.5 Trend comparison for the three protein lists with the aggregate trend of the learning set list of 268 proteins. 143

C.6 Distributional analysis of the query residues for the learning set list of 268 proteins. 144

Chapter 1

Introduction

The proper functioning of life is a gift of the dominant role played by proteins in our body. Therefore, one of the most important tasks is to understand these polymerized combinations of amino acids that form the essence of life - the proteins. The function of proteins is governed by their structure, which in turn is a direct consequence of amino acid sequence. In most of the cases, protein function has been more easily and widely studied than its detailed structure. A proper understanding of both protein structure and function can help us in drug design. New and improved biomaterials like enzymes, protein computers and biochemical machines can be synthesized with this knowledge.[1] Also, a proper understanding of inter-atomic interactions of proteins is very important in the prediction of protein structure and also in the design of novel macromolecules.[2] For the development of high throughput screening techniques, it is important to decipher protein structure in the simplest possible way.

Broadly speaking, proteins have two types of residues - core residues and surface accessible residues. The residues that form the core of the protein are engaged in maintaining the structure and stability of the protein and the residues that are accessible to the solvent are responsible for the various activities of the proteins like, ligand binding and catalytic activities. In order to select a protein sequence as a drug target, the identification of its corresponding amino acids as buried or surface accessible becomes relevant.[3] Moreover, surface accessibility predictions in conjunction with secondary structure information can be exploited to devise computational methods for the prediction of protein tertiary structure. Hence, to develop a software that quickly and effectively predicts protein residue surface accessibility and also the tertiary structure of the protein, this thesis presents the first step forward- the characterization of surface accessible residues and related effects.

1.1 Existing Approaches

Two fundamental approaches, experimental and computational, can be employed for the identification and characterization of the accessible surface area and hence the residues associated with them. The experimental methods are very accurate and reliable in the sense that these methods are associated with close to zero false positives and a very tight confidence interval. But, they are time and labor consuming, relatively few protein structures can be determined with these methods within a given time frame. Also, the total cost of structure determination or evaluation for annotation by these traditional methods is more compared to the computational methods. Hence, the computational methods are well suited for any high throughput screening methods as they are time, labor and cost effective.

Although the computational methods are time efficient and cost effective, and are capable of predicting the structure of huge data-sets, they are not reliable and high resolution enough to provide a realistic replacement of the experimental methods. However, they are still useful in proteomics to get incomplete or low-resolution structure information for structural comparisons[4] and evaluation of protein-protein and ligand-protein interactions when screening large number of sequences.

1.1.1 Experimental Methods

The experimental methods used for protein structure determination include X-ray diffraction studies, electron diffraction methods, electron microscopy and nuclear magnetic resonance (NMR) technique.[5] High resolution three-dimensional protein structures can be obtained from X-ray diffraction studies, but this method has an inherent limitation due to the difficulties in obtaining protein crystals.[6] With methods like electron diffraction and electron microscopy, medium to low resolution outputs can be obtained respectively.[5] Here resolution proves a problem and restricts such protein structures to some specific applications only. Although peak resolution is sometimes problematic for protein structure determination, NMR technique is the best available method as it has the potential to determine protein structure with high resolution in its native environment.[7]

1.1.2 Protein Structure Modeling Approaches

The interdependence of protein sequence structure and function suggests that in order to do functional annotations, calculation of three dimensional model of a protein becomes inevitable or necessary.[8] The accuracy of functional annotations relies on the modeling approach taken for its predictions. The computational approaches for protein structure modeling include comparative modeling, threading and ab initio or de novo methods.

1.1.2.1 Comparative Modeling

Comparative model building includes either sequential or simultaneous modeling of the core of the protein, loops and side chains.[5,9] The comparative modeling approach taps into the sequence-structure relationship for its predictions. A small change in protein sequence manifests itself in a small change in 3D structure. Hence, in order to predict a useful model, detectable sequence similarity and presence of correct alignment between template and target protein is a must. Structures modeled by this method are comparable in resolution to either low-resolution X-ray structures or medium resolution NMR solution structures.[5] Comparative modeling does not typically allow for detailed structures regarding protein function or related drug design. The challenges faced by this method are posed by the limited accuracy of sequence-structure alignment and difficulty in loops and side chain modeling. An absence of techniques for accurate modeling of rigid body shifts and distortions, absence of scoring functions for the measurement of model quality and local error detection also limit the scope of this method.

1.1.2.2 Threading

In this method, sequences are threaded onto all known folds and evaluated on the basis of goodness of fit.[10–12] The scores are calculated by assignment and alignment of the threaded sequence to each of the structures in a library of all known folds. Although threading methods are more sensitive than sequence comparison methods,[10] there are certain inherent limitations to this approach.[13] It is difficult to get the exact fold match for some structural analogues and remote homologues.[11,12] Predictions using these techniques give a low confidence interval with the chance of an incorrect output. Also, after a correct fold recognition, the accuracy of threading alignment is not very high $(60 - 90\%)$ for proteins with $< 30\%$ sequence identity.[14] Only a limited percentage of protein families can be predicted successfully using this approach.[13] Also, a need for new and improved threading energy functions, algorithms, evaluation and refinement techniques has become more evident.

1.1.2.3 *Ab Initio or de Novo*

de Novo methods predict the structure directly from protein sequence without relying on any similarity between the protein sequence under question and any of the known protein structures.[5] *Ab initio*[4,15] assumes that the native structure of a protein is the global free energy minimum and predicts the protein structure by solving a minimization problem of a unified physical energy function.[9] Hence, the dependence of these models on the laws of physics poses a limitation, because either the physical models or the level of computational efficiency is not adequate for the prediction requirements.[13] It is computationally intensive, limited to smaller proteins $(< 100 - 150$ residues) and less accurate than the template based methods. Also, it is useful in cases when either the template does not exist or is marked by an alignment problem. It is also useful in determination of non-homologous loop regions.[16]

1.1.3 Sequence Homology Methods

Sequence homology method is a computational method for modeling the structure of a protein based on its sequence similarity to one or more other proteins of known structure. This approach is involved in partial protein structure evaluation and deals with structural annotations. Although the sensitivity of homology methods is limited, the relative ease and simplicity makes this method a versatile tool for partial structure predictions. This method is based on protein homology and makes predictions exclusively from sequence, which are expected to be more detailed and reliable. On account of being most detailed and accurate, homology-based or comparative modeling approaches are the methods of choice for protein structure modeling. Also, many biological processes can only be understood if explained at the amino-sequence level.[17] The presence of at least one known protein structure, that is "recognizably related" to the sequence that is being modeled is a must in case of comparative modeling. However, for sequence homology-based models, no such pre-requisites are needed.

Due to the interdependent relationship between protein sequence, structure and function, structural annotations can be made by comparing statistical parameters in an aligned set or subset of protein residues.[18] For studying a larger database of homologues that are functionally related, this is the method of choice. A direct relation between the evolutionary changes and the distribution of any particular residue type at any given aligned residue position can be attributed to the constraints imposed by the function of that residue. Statistically significant signals are expected in case of co-evolution of residues and hence can be related easily with the functional signals. Basic Local Alignment Search Tool (BLAST) and Fast Alignment (FASTA) are the two basic tools for sequence homology methods.[19]

The goal of this work was to provide some structural annotations, that is, classify residues as buried or surface accessible, as well as provide some insight into the protein folding ‘problem’ in the easiest, time efficient manner. Hence, the sequence homolgy based method that takes care of the “long range communications”, that are “crucial for biological functions”, between distant protein homologues, was selected as the method for this project.[18] The merit of this method lies in its fast and easy calculations that can be used to screen thousands of proteins and loops and also in its potential to incorporate experimental data.

1.2 Important Terms

1.2.1 Packing Density

Starting from the efforts of Richards[20] in the calculation of protein packing density with the help of Voronoi polyhedra, a number of other approaches like, the Delaunay tesselation,[21] the coarse grained scale[22] and the mass size exponent[23] were among the many methods proposed for the same purpose. In this thesis, the relative packing of protein residues, calculated from the crystallographic coordinate data, has been termed packing density. Packing density is calculated by taking into account all the residues that fall within an appropriate radius from the amino acid in question. It is defined as the number of C_α atoms falling in the radius of $9\AA$, around the residue of interest. The packing density at any residue position K is given by equation 1.1.

$$D_K R = N_K(persphereofradiusR) \tag{1.1}$$

where, $D_K R$ is the number N of C_α carbons found within a radius of R from the C_α position of residue K.

The distance between any two residues is calculated by equation 1.2.

$$dist(i,j) = \sqrt{(x(i)-x(j))^2 + (y(i)-y(j))^2 + (z(i)-z(j))^2} \tag{1.2}$$

where, x, y and z are the C_α coordinates at that position.

It provides an estimate of how well an amino acid is surrounded by the neighboring residues. It gives a crude estimate of the distance of the residues from the core of the protein. Also, it provides an estimate of flexibility at a protein sequence position. Stability of the protein relies on the packing of the core of the protein.[2] High packing density corresponds to the stability of the overall protein fold and does not necessarily translate to an essentially closed or compact global protein structure.[24]

Flexibility is an inverse measure of packing density.[25] It indicates the amount of motion allowed at that residue position. Flexibility allows a protein to bind to more than one type of substrate. The greater the flexibility at a particular residue position, the more mutable are the corresponding residues.

1.2.2 Sequence Entropy

Entropy is a measure of total randomness in a system. Sequence entropy, in case of proteins, provides a measure of partitioning of a particular type of residue at any aligned residue position. At any sequence position k, it is defined by the equation 1.3

$$S_k = -\sum_{j=1,20} P_{jk} * \log_2 P_{jk} \tag{1.3}$$

where, the probability P_{jk} at any sequence position k is obtained from the frequency of an amino acid type j at sequence position k for all the aligned residues.[26]

Sequence entropy calculations is based on information theory and hence it is sometimes referred to as Shannon entropy. Sequence entropy can be correlated to configurational entropy.[27] Also, a correlation between sequence entropy and inverse packing density[26] indicates the presence of two major regions that can serve as a basis for surface accessibility predictions.

1.2.3 Surface Accessibility

The identification of protein residues of the monoclonal bodies as surface accessible or buried can help in antibody development.[3,28] Also, once the surface accessiblity of a residue is understood, protein tertiary contacts, boundary between structural domains, intramolecular rearrangements and protein hydration sites can be predicted with much ease.[29]

A number of methods, involving residue substitution matrices, Bayesian statistics, neuronal networks and the residue level characterization of the surface exposed residues, have been proposed. These methods either characterize the residues according to the binary classification (solvent exposed or buried, associated with $70-75\%$ prediction success), ternary classification (buried, partially or completely surface exposed, associated with 55% prediction success), or more complex classification (associated with $20-25\%$ prediction success) that categorizes the residues into approximately 10 groups.[29] Machine learning approaches, like the support vector machine and the related algorithms are known to provide better percentage accuracy ($80-90\%$) in such predictions.[30]

NACCESS is one of the many tools available for protein surface accessibility predictions that deal with the protein surface area at the atomic level.[31] It calculates the accessible surface for each atom and also provides an average surface accessibility value per protein residue by rolling a probe of given size around a protein surface. In this work, relative surface accessibility values provided by the NACCESS program have been utilized for the estimation of surface accessibility of the individual protein residues. Since the binary classification predicts surface accessibility more accurately than any other currently used methods and also, since the calculations based on aligned homologous sequences are known to further increase the prediction accuracy,[29] for this work sequence homology-based methods have been utilized for the binary prediction of a set of 268 proteins.

1.3 Definitions

- Bit score - Bit score represents the statistical significance of alignments. Higher the bit score, the more similar are the aligned sequences.

- BLASTP- Basic Local Alignment Search Tool for Proteins is a tool that utilizes heuristic method for aligning all the query residues to the subject residues. It identifies first the local regions of similarity followed by an identification of global alignment. It compares protein sequence with a protein database.

- E-value - Expectation value or the E-value provides the likelihood of chance similarity between the query and the subject sequence. The lower the E-value, the more similar the sequences are within a more tight confidence interval.

- Fraction gaps - Fraction of gaps is defined as the total number of gaps divided by the total number of aligned residues at that aligned residue position. In the 'gaps-excluded' cases it is essentially a ratio between the total number of gaps and the total number of residues at any aligned residue position.

$$Fraction\ gaps_i = \frac{Number_{gaps_i}}{Total\ Number\ of\ Residues_i} \tag{1.4}$$

  Where, $Number_{Gaps_i}$ is the number of Gaps (-) at sequence position i.

- Fraction non-strongly hydrophobic - Fraction of non-strongly hydrophobic residues is defined as the total number of residues that are not strongly hydrophobic divided by the total number of aligned residues at that aligned residue position.

$$Fraction\ non-strongly\ hydrophobic_i = 1-Fraction\ strongly\ hydrophobic_i \tag{1.5}$$

- Fraction small residues - Fraction of small residues is defined as the total number of residues that are small divided by the total number of aligned residues at that aligned residue position.

$$Fraction\ small\ residues_i = \frac{Number_{SR_i}}{Total\ Number\ of\ Residues_i} \tag{1.6}$$

  Where, $Number_{SR_i}$ is the number of small residues (A and G) at sequence position i.

- Fraction strongly hydrophobic - Fraction of strongly hydrophobic residues is defined as the total number of residues that are strongly hydrophobic divided by the total number of aligned residues at that aligned residue position.

$$Fraction\ strongly\ hydrophobic_i = \frac{Number_{SHP_i}}{Total\ Number\ of\ Residues_i} \tag{1.7}$$

  Where, $Number_{SHP_i}$ is the number of strongly hydrophobic residues (VILFYMW)[32] at sequence position i.

- Heterodimers - Proteins that consist of two non-identical pair of polypeptides and share sequence identity $< 90 - 95\%$.[33]

- Homodimers - Proteins which have two chains of identically sequenced polypeptides that share at least $90 - 95\%$ sequence identity.[34]

- Homology- A similarity between protein sequences attributed to common ancestral origin.[19]

- Monomers - Proteins with only one polypeptide chain.

- Packing density - Packing density is defined as the total number of residues that surround the residue of interest and is calculated by counting the total number of residues that fall inside a radius of $9\AA$ from the residue of interest.[35]

- Percentage sequence identity - It is defined as the number of identical residues divided by the number of matched residues, where gaps are not taken into account.

- RSA - Relative Surface Accessibility can be defined as the relative solvent accessibility of the protein residues.

- Sequence entropy - Sequence entropy is a measure of variability of the query protein sequence that is calculated by summing the total variability at each aligned residue position.

1.4 Overview

In this thesis, I have summarized the work on the selection of a set of homology-based thresholds that would be apt for the surface accessibility prediction of protein residues according to the binary classification as buried or surface accessible. Sequence homology-based approach was utilized for this analysis. First of all a diverse set of proteins that satisfy some basic standards, as discussed in Chapter 2, was compiled. All the protein sequences were aligned to similar protein sequences with the help of BLASTP. From the aligned residues, entropy and all the fractional parameters were calculated. For each of the residues, packing density was calculated and aligned with the homology based parameters for the aggregate learning set list of protein residues. The aggregate correlation plots of homology-based parameters as a function of inverse density were then analyzed. The aggregate trends of monomeric,[36] homodimeric[34] and heterodimeric[33] proteins were compared to each other and also to the learning set list of proteins (discussed in Chapter 3). Then the frequency distribution plots of the various homology-based parameters and the aggregate correlation plots of the learning set list of proteins, were evaluated in light of filter threshold selection for the binary classification of residues as buried or surface accessible. Conservative filter thresholds were selected for three homology-based parameters- entropy, fraction small residues and fraction strongly hydrophobic. They are designed to be applied to the aggregate aligned protein set (learning set). The two resulting frequency distributions and cumulative frequency distributions were then analyzed for their surface accessibility prediction accuracy.

1.5 Organization of the Thesis

This thesis is organized into six chapters - Introduction, Methods, Results, Discussion, Conclusion and Future Studies. In the first chapter (Introduction), the importance and goal of this work has been discussed. The prevalent approaches to accomplish the objective and the advantage of using sequence homology methods over all the other methods have been explained to a certain extent. Here, all the important terms have been defined and the key phrases have been discussed in some detail. The second chapter (Methods), focuses on the details of the various techniques applied to this work in order to produce a fruitful outcome, the prediction of residues as surface accessible or buried with a good level of accuracy. This chapter deals with the compilation of the various protein lists, generation of an aggregate aligned set of protein residues with calculations of entropy, density and fractional parameters, generation of various aggregate correlation plots and generation of various frequency distribution plots. All the tables and figures are embedded in the text of the third chapter - Results. Here, all the observations and inferences related to the various plots and frequency distributions have been discussed. Chapter 4 (Discussion) explains all the results obtained from the various analyses and their relevance in light of the objective of this work. Here other observations that provide some insight into the physical aspect of the protein have also been discussed briefly. Other approaches to surface accessibility prediction have also been compared here with the approach undertaken for this work. The limitations of this approach has also been acknowledged here. The fifth chapter (Conclusion) concludes all the methods and results and provides reasonable homology-based filter thresholds for the prediction of surface accessible residues. The suggested future work has been enumerated in Chapter 6, Future Studies.

Chapter 2

Methods

In order to characterize structural features of proteins, sequence homology-based parameters and their relation with C_α packing density were explored. A structurally diverse set of protein lists was prepared followed by calculation of density and homology-based parameters at each protein residue position. Then from the resulting data, various correlation plots and frequency distribution plots were generated for further interpretation and analysis. This chapter explains these methods in detail.

2.1 Protein Set Preparation

A subset of 281 proteins that were specifically studied by Guharoy and Chakrabarti[37] was obtained by culling the 281 list with PISCES[38] according to a set of parameters. The relationships between the various entries were evaluated on the basis of PSI-BLAST and CE structural alignments. The 281 protein set was culled for PDB chain identifiers that share sequence percentage identity of $\leq 25\%$, have a structural resolution of $0.0 - 2.5\mathring{A}$, R-factor ≤ 0.3 and sequence length $40 - 10,000$. Protein chains with C_α only entries were eliminated. Proteins whose structure was determined by X-ray crystallography methods were only selected. This set of user defined culling criteria was standardized for culling all the other lists used in this work. A set of 215 proteins so obtained was named Chack05 list.

For the preparation of learning set list, the set of 281 proteins mentioned earlier, was combined with a structurally diverse set of well characterized 130 query proteins[26] with known X-ray 3D structures. The resulting composite set of 408 proteins were then culled according to the standard criteria mentioned above with the help of the culling server, PISCES.[38] The set of 268 protein chain list specified with their PDB chain identifiers was named the learning set list of 268 proteins. The learning set list consists of three types of proteins: monomeric proteins,[36] homodimeric proteins[34] and heterodimeric proteins.[33] In order to develop a better understanding of their respective contributions to the aggregate behavior of the learning set list of 268 proteins, monomers, homodimers and heterodimers were evaluated independently. The protein lists of 75 monomers, 106 homodimers and 50 heterodimers were prepared by culling the lists of 103 monomeric proteins,[36] 122 homodimeric proteins[34] and 70 heterodimeric proteins[33] respectively, in PISCES, according to the standard criteria determined as above.

2.2 Residue Packing Density

Residue packing density is the packing density of a protein residue in its native state. It is a measure of protein compactness. It is obtained by a coarse grained approach that takes into account the X-ray determined C_α coordinates of the query protein for its computation.

For the calculation of protein residue packing density, mmcif files for all the protein lists obtained as above were downloaded from RCSB Protein Database (PDB).[39] The perl ftp script *ftp-script-1.pl* (see Appendix A) was used for obtaining all the mmCIF files. The atom co-ordinate information of the proteins that is obtained from X-ray crystallographic studies are compiled in the mmCIF files.[40] All the *.Z* files obtained as above were *gunzipped* on the Cluster (Unix environment) provided by the Meteorology Department of San Jose State University. The resulting *.CIF* files, written in mmCIF file format, served as the source of the residue coordinate information. With the help of a perl program - *Cif2Den.pl* (see Appendix A), that was adapted from Yeh,[35] the C_α packing density of the protein was calculated at each residue position. First of all, the C_α co-ordinates were extracted at each residue position. The distance between any two residues was then calculated by equation 2.1.

$$dist(i,j) = \sqrt{(x(i) - x(j))^2 + (y(i) - y(j))^2 + (z(i) - z(j))^2} \qquad (2.1)$$

where, x, y and z are the C_α coordinates at that position. The number of C_α atoms falling in the radius of $9\mathring{A}$, around the residue of interest was counted. The packing density at that residue position was then calculated by using the equation 2.2.

$$D_K R = N_K(persphereofradiusR) \qquad (2.2)$$

where, $D_K R$ is the number N of C_α carbons found within a radius of R from the C_α position of residue K.

The program *Chainselectivecif2den.pl* (see Appendix A) was used for the packing density calculations of all the protein lists. Packing density equal to 0 was assigned to the unknown residues like *'X'* and a density value equal to *'NA'* was assigned to the residues whose coordinate information was not available in the mmCIF files.

2.3 Sequence Variability

The importance of a protein residue is depicted in its evolutionary conservation. Changes at specific position of protein that preserves physico-chemical properties of the original residue is called sequence conservation. Sequence variability, an inverse measure of sequence conservation, provides an estimate of the role played by the residue in maintaining the structure and function of the protein through the years of evolution. Although a number of other conservation scoring strategies are present, in this work, the sequence entropy at any protein residue position has been calculated either by calculating the Shannon entropy (information entropy or sequence entropy) from the alignments generated from BLASTP or from multiple sequence alignments provided by the HSSP database.

2.3.1 Calculation of Sequence Entropy

Sequence entropy or Shannon entropy is the metric of sequence variability. It is a measure of disorder or randomness in a system[35] and at any sequence position k, it is defined by the equation 2.3.

$$S_k = -\sum_{j=1,20} P_{jk} * \log_2 P_{jk} \tag{2.3}$$

where, the probability P_{jk} at any sequence position k is obtained from the frequency of an amino acid type j at sequence position k for all the aligned residues.[26]

Shannon entropy correlates to thermodynamic entropy.[27] Entropy is expected to play a significant role in deciphering the relation between protein stability and function.[26] For the calculation of sequence entropy, the query protein is aligned to other subject protein sequences present in the database. The resulting alignments are then parsed and entropy values are calculated at each residue position of the query protein.

The sequence alignments for all the protein lists were generated by Basic Local Alignment Search Tool (BLASTP $2.2.18+$), that is provided by the National Center for Biotechnology Information (NCBI). BLASTP used for this work searches all the non-redundant protein databases like, Genbank, CDS translations, PDB, SwissProt, PIR and PRF but excludes all the environmental samples from WGS project. It uses BLOSUM62 matrix and default gap penalties for each mutational insertion or deletion. The FASTA formatted query sequences were submitted for the BLASTP alignments at NCBI. For each protein sequence, a maximum of $10,000$ aligned sequences were generated for this work. An existence of 11 and extension of 1 was set as the cost to create and extend a gap in an alignment. In order to compensate for the amino acid composition of a sequence, conditional compositional score matrix adjustment method was used.[19] All the alignments were generated between June

1st and September 1st 2008, when the total number of sequences in database was approximately $6,923,879$. The aligned residues were extracted from the BLASTP results by using a code written in perl- *bst2entMOD2.pl* (see Appendix A). A total of 75891, 20075, 29427 and 9229 alignments were generated for the learning set list, the monomeric list, the homodimeric list and the heterodimeric list, respectively.

For the calculation of sequence entropy, the BLASTP alignments with bit scores equal to 40% of the highest bit score obtained in a set of alignments, were only considered for entropy calculations. The resulting 40% sequence entropy at any protein residue position is Shannon entropy calculated from the BLASTP alignment results at that position. Sequence entropy was calculated by a perl code *bst2entMOD2.pl* (see Appendix A) that uses equation 2.3 for its entropy calculations.

The 40% sequence entropy so obtained is based on the bit score that qualifies the sequence for the goodness of an alignment. The advantage of using bit score cutoffs is, that it is normalized and can be compared against results derived from other substitution matrices.[35] Among the other bit score cutoffs, 40% cutoff has been reported to provide the most relevant set of alignments[25,35] because it balances between homology and the diversity of sequence variability. A number of other cutoffs like, an expect cutoff of 0.001 (see Appendix Figure B.1) and an identity cutoff of 25% (see Appendix Figure B.1) were also tested for obtaining quality alignments. Introduction of gaps as the 21st term in the entropy calculations was also tested (see Appendix Figure B.2, B.3 and B.4). The minimum value of non-deleted homologs for entropy calculations was set to 1, so that a “divided by zero” error could be prevented. Entropy for all the sequence positions with less than this minimum value were assigned a ‘-1’ entropy value.

2.3.2 6-point Entropy

The 6-point entropy[41] at any query residue position was calculated from the alignments generated for 40% entropy calculations. In this case, the 20 amino acids were categorized into six clusters according to their classification as aliphatic (AVLIMC), aromatic (FWYH), polar (STNQ), positive (KR), negative (DE)and special (GP). The 6-point entropy was calculated by the equation 2.4.

$$S_l = -\sum_{i=1}^{6} p_i * \log_2 p_i \tag{2.4}$$

where, p_i represents the frequency of each of the six i classes at any given query residue position. A low value of 6-point entropy, translates to evolutionary conservation at that point. For the calculation of 6-point entropy, a perl code- *Radhika-*6*pointentropy.pl* was used (see Appendix A).

2.3.3 Calculation of HSSP Entropy

HSSP entropy are the entropy values obtained directly from the HSSP files provided by Homology derived structures of proteins (HSSP) database.[42–44] Here, the query sequence of known structure is aligned to other homologous sequences. The homology between sequences is determined on the basis of a homology threshold curve that is plotted after a detailed investigation of sequence similarity, structure similarity and the protein alignment length. The HSSP files contains structure based multiple sequence alignments, sequence variability at each position and sequence profile. The likelihood of the aligned sequences to share the same three-dimensional structure is very high. Hence, information embedded in these files can be used to explain the role of conserved residues in protein structure as well as in deriving patterns for structure prediction. The residue conservation scores compiled in the HSSP files,[44] are expressed as Shannon entropy given by equation 2.5.

$$V_{schneider} = -\sum_{i}^{j} p_i * \ln p_i \tag{2.5}$$

where, $j = 20$ that represents 20 amino acids and p_i is the fractional frequency of amino acid of type i. Entropy calculated in this fashion lies between $0 \leq V_{schneider} \leq 1$.

The 254 HSSP files belonging to the learning set list of 268 proteins were downloaded from the HSSP database in February 2008 with the help of the perl code-*ftp-scriptHSSP.pl* (see Appendix A). Some 26 protein chains belonging to 12 proteins (11 of which belong to the Chack05 list and one belongs to the 130 list of proteins) have multiple entries in the learning set list of proteins. Hence, fewer number of files (254) than 268 were downloaded as only one HSSP file per protein can be obtained from the HSSP database. The total number of multiple sequence alignments for the 254 HSSP protein list was 235810. The enlisted entropy values at each residue position was extracted programmatically.

2.4 Fractional Calculations

In addition to sequence entropy, other sequence-homology based parameters like fraction of residues that are strongly hydrophobic, fraction of residues that are small, fraction of residues that are non-strongly hydrophobic and fraction of residues that are gaps were also calculated. The aligned residues, obtained as in section 2.3.1, were used for these calculations. For the calculation of strongly hydrophobic fraction at any position i, all the aligned residues that were strongly hydrophobic at position i were counted and then divided by the total alignment length at that i position (equation 2.6). All the other fractional parameters were similarly calculated (equation 2.7, 2.8 and 2.9).

$$Fraction\ strongly\ hydrophobic_i = \frac{Number_{SHP_i}}{Total\ Number\ of\ Residues_i} \tag{2.6}$$

Where, $Number_{SHP_i}$ is the number of strongly hydrophobic residues (VILFYMW)[32] at sequence position i.

$$Fraction\ non\ strongly\ hydrophobic_i = 1 - Fraction\ strongly\ hydrophobic_i \quad (2.7)$$

$$Fraction\ small\ residues_i = \frac{Number_{SR_i}}{Total\ Number\ of\ Residues_i} \quad (2.8)$$

Where, $Number_{SR_i}$ is the number of small residues (AG) at sequence position i.

$$Fraction\ gaps_i = \frac{Number_{gaps_i}}{Total\ Number\ of\ Residues_i} \quad (2.9)$$

Where, $Number_{Gaps_i}$ is the number of Gaps (-) at sequence position i.

The fraction gaps is not a fraction but a ratio between the total number of mutational insertions and deletions and the total number of residues present at that query position. All the fractional values were calculated and aligned with the entropy values with the help of a perl code, *extract_fractanalysis_entropy_aggr.pl* (see Appendix A). The resulting *.FRAC* files were then further processed.

2.5 PDB-FASTA Reconciliation

The residue position numbers mentioned in the *.PDB* files obtained from Protein Data Bank do not necessarily match with the residue position numbers of the same protein obtained from FASTA format numberings.[45] Hence, for the proper alignment of entropy, fractional parameters and density, the position numbers of the density (*.DEN*) files were matched programmatically with that of the position numbers of the entropy (*.ENT*) and fraction (.FRACT) files. The perl code, *extract_individualfractentropy_density_aggr.pl* was used for this work (see Appendix A).

In case of a few proteins the first few residues were found to be eliminated in the alignments generated by BLASTP. In those cases, the alignments were found to shift one or two places to the left thereby assigning false residue numbers to the residues, for example, a residue number of 1 was assigned to residue 2 and so on. These disparities were checked manually, and the *.FRACT* files were modified accordingly for the proper FASTA-PDB alignment.

2.6 Aggregate Analysis

For the aggregate analysis of protein lists, all the individually aligned protein files obtained as above were further processed. The average value of homology based parameters at each density position was obtained by the method of single averaging. All the parameter values at a particular density position were averaged to obtain the single average value of that homology based parameter at that packing density. The perl code, *calculate_aggr_per_protein.pl*, was used for this calculation (see Appendix A). Then all the single aggregate values belonging to the complete protein list were compiled into one single file with the help of the perl code, *double_aggr_forPlot.pl* (see Appendix A). The file so obtained was then processed in Microsoft Excel 2007. Double average value of the parameters were obtained by averaging all the single average values compiled as above, at a particular density position for the complete protein list.

2.6.1 Correlation Plots

The aggregate file obtained from the code double_aggr_forPlot.pl (see section 2.6) was then processed in MS Excel. The aggregate value of all the parameters at each density position was calculated. For the computation of double average value, all the homology-based parameters at any particular density was averaged. The resulting value was the double average value at that density position. Different correlation plots were obtained by plotting various homology-based parameters against inverse density.

2.7 Frequency Distributions

The entropy-fractional parameter alignment files, that is, the *.FRACT* files (see subsection 2.4) for the complete list were aligned to the respective density values at each residue position and compiled into one single file. This work was made possible by the perl code, extract_fractentropy_density_aggr.pl (see Appendix A). The file so obtained was processed by Microsoft Excel 2007. First of all, the density position was sorted from smallest to largest value and from A to Z with the 'expand selection' option. All the rows that contained density values equal to 0, or density values equal to 'NA' were eliminated from the file. Density equal to 0 represents the density value of an unknown protein residue denoted by 'X'. Density equal to 'NA' represents the sequence position for which the co-ordinate information is not available. All the residue positions, flagged by an entropy value of -1 were also removed before obtaining the frequency distributions. The resulting *.XLS* file was then used to obtain the various frequency distributions.

2.8 Characterization of Protein Lists

Each of the protein lists was characterized by generating frequency distribution graphs for a set of parameters. The frequency of query proteins with respect to packing density was obtained from the final *.XLS* file as obtained in subsection 2.7. The frequency of query proteins versus number of alignments histogram was obtained with the help of the perl code *listNoAlignments.pl* (see Appendix A) and MS Excel 2007. The frequency of query proteins to length of query proteins histogram was obtained with the help of the perl code *No_of_res_count.pl* (see Appendix A) and MS excel 2007. The frequency of subject proteins at BLAST bit score was obtained with the help of the perl code *Bitscorelistno_ofsubject.pl* (see Appendix A).

Chapter 3

Results

In this chapter, the results obtained by the analysis of the four protein lists (see Chapter 2), namely, learning set list (268 protein chains), monomeric list (75 proteins), homodimeric list (106 proteins) and heterodimeric list (50 protein chains) have been summarized. The protein lists have been characterized by various frequency distributions. The aggregate behavior of the lists are presented by means of correlation plots of the various homology-based parameters. Aggregate trends presented by each of the homology-based parameters has been studied both separately and in conjunction with each other. In order to aid a detailed analysis, frequency distribution plots based on density and the value of homology-based parameters have been reported here separately. The homology-based filter thresholds for the prediction of surface accessibility have also been evaluated here.

3.1 Characterization of Protein Lists

In this section, all the four protein lists have been characterized on the basis of their mode of structure determination, resolution, R-factor, Free R value, protein length and their alignment length. The frequency distribution plots of the query proteins with respect to different parameters has also been summarized here for the four protein lists.

3.1.1 Characterization of Learning Set List

Table 3.1 summarizes the learning set list of 268 proteins and the PISCES culling parameters like protein chain name, the experimental method of structure determination, resolution, R-factor and free R value of the constituent proteins (see Chapter 2).

Table 3.1: Learning set list of 268 proteins. A composite list of 130 proteins and chack05 list of proteins were culled by PISCES according to individual protein chains of each PDB ID. Each protein chain with sequence percentage identity $>$ 25% was rejected. All these protein chains have a resolution of $0.0 - 2.5\mathring{A}$, R-factor ≤ 0.3 and sequence length between $40 - 10000$.

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
12AS	A	XRAY	2.2	0.16	0.29
13PK	A	XRAY	2.5	0.22	0.29
1A1I	A	XRAY	1.6	0.19	0.22
1A2K	A	XRAY	2.5	0.21	0.27
1A32	A	XRAY	2.1	0.21	0.32
1A48	A	XRAY	1.9	0.15	1
1A4I	A	XRAY	1.5	0.2	0.23
1A4U	A	XRAY	1.92	0.2	0.24
1A6Q	A	XRAY	2	0.21	1
1AA7	A	XRAY	2.08	0.21	0.28
1ADD	A	XRAY	2.4	0.18	1
1ADE	A	XRAY	2	0.2	1
1AF3	A	XRAY	2.5	0.23	0.27
1AFW	A	XRAY	1.8	0.19	0.24
1AG9	A	XRAY	1.8	0.2	0.25
1AH7	A	XRAY	1.5	0.2	0.23
1AJS	A	XRAY	1.6	0.17	1
1AK0	A	XRAY	1.8	0.21	0.23
1AK4	C	XRAY	2.36	0.24	0.31
1AKO	A	XRAY	1.7	0.17	0.2
1AL8	A	XRAY	2.2	0.19	0.25
1AMK	A	XRAY	1.83	0.11	1
1AMP	A	XRAY	1.8	0.16	1
1AMU	A	XRAY	1.9	0.21	0.25
1AN9	A	XRAY	2.5	0.2	0.26
1AOB	A	XRAY	2.1	0.19	0.24
1AOR	A	XRAY	2.3	0.15	1
1AQ0	A	XRAY	2	0.17	0.21
1AQ6	A	XRAY	1.95	0.19	0.25
1ATL	A	XRAY	1.8	0.16	1
1AUO	A	XRAY	1.8	0.21	0.27
1AVW	B	XRAY	1.75	0.19	0.21
1AW5	A	XRAY	2.3	0.2	0.27
1AW7	A	XRAY	1.95	0.18	1
1AW9	A	XRAY	2.2	0.2	1

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
1AYL	A	XRAY	1.8	0.2	0.23
1AYX	A	XRAY	1.7	0.15	0.18
1AZI	A	XRAY	2	0.17	1
1B3A	A	XRAY	1.6	0.17	0.24
1B5E	A	XRAY	1.6	0.19	0.21
1B67	A	XRAY	1.48	0.19	0.27
1B8A	A	XRAY	1.9	0.17	0.2
1B8J	A	XRAY	1.9	0.18	0.2
1BA3	A	XRAY	2.2	0.2	0.24
1BAM	A	XRAY	1.95	0.19	1
1BBH	A	XRAY	1.8	0.18	1
1BD0	A	XRAY	1.6	0.24	0.27
1BEA	A	XRAY	1.95	0.2	0.29
1BF2	A	XRAY	2	0.16	0.21
1BFD	A	XRAY	1.6	0.15	0.19
1BG0	A	XRAY	1.86	0.2	0.22
1BIA	A	XRAY	2.3	0.19	1
1BIN	A	XRAY	2.2	0.2	0.3
1BIQ	A	XRAY	2.05	0.19	0.26
1BIS	A	XRAY	1.95	0.2	0.26
1BJW	A	XRAY	1.8	0.21	0.27
1BLZ	A	XRAY	1.45	0.2	0.22
1BMD	A	XRAY	1.9	0.15	1
1BN6	A	XRAY	1.5	0.17	0.17
1BO6	A	XRAY	2.1	0.21	0.25
1BRS	A	XRAY	2	0.17	1
1BRS	D	XRAY	2	0.17	1
1BRW	A	XRAY	2.1	0.23	0.28
1BSL	A	XRAY	1.95	0.19	1
1BT3	A	XRAY	2.5	0.17	0.25
1BUL	A	XRAY	1.89	0.21	0.26
1BUO	A	XRAY	1.9	0.21	0.25
1BXG	A	XRAY	2.3	0.17	1
1BXK	A	XRAY	1.9	0.2	1
1BXQ	A	XRAY	1.41	0.14	0.18
1BYO	A	XRAY	2	0.19	0.23
1C02	A	XRAY	1.8	0.2	0.25
1CB0	A	XRAY	1.7	0.18	0.2
1CDC	A	XRAY	2	0.19	1
1CEX	A	XRAY	1	0.09	0.12
1CG2	A	XRAY	2.5	0.2	0.22
1CHM	A	XRAY	1.9	0.18	1

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
1CJX	A	XRAY	2.4	0.22	0.28
1CKI	A	XRAY	2.3	0.19	0.28
1CMB	A	XRAY	1.8	0.19	1
1CNZ	A	XRAY	1.76	0.2	0.26
1COZ	A	XRAY	2	0.2	0.26
1CQX	A	XRAY	1.75	0.18	0.21
1CRC	A	XRAY	2.08	0.18	1
1CRM	A	XRAY	2	0.18	1
1CRZ	A	XRAY	1.95	0.19	0.24
1CSE	E	XRAY	1.2	0.18	1
1CSE	I	XRAY	1.2	0.18	1
1CSH	A	XRAY	1.65	0.16	1
1CTT	A	XRAY	2.2	0.19	1
1CZJ	A	XRAY	2.16	0.2	0.26
1DAA	A	XRAY	1.94	0.18	1
1DAN	L	XRAY	2	0.19	0.22
1DAN	T	XRAY	2	0.19	0.22
1DAN	U	XRAY	2	0.19	0.22
1DCS	A	XRAY	1.3	0.13	0.15
1DFJ	I	XRAY	2.5	0.19	1
1DHK	A	XRAY	1.85	0.18	0.22
1DHK	B	XRAY	1.85	0.18	0.22
1DHS	A	XRAY	2.2	0.15	0.24
1DHT	A	XRAY	2.24	0.19	0.28
1DIN	A	XRAY	1.8	0.15	1
1DMR	A	XRAY	1.82	0.15	0.18
1DOR	A	XRAY	2	0.17	0.21
1DPG	A	XRAY	2	0.21	0.26
1DQS	A	XRAY	1.8	0.17	0.22
1DYS	A	XRAY	1.6	0.18	0.24
1E1K	A	XRAY	1.95	0.18	0.23
1E5M	A	XRAY	1.54	0.17	0.2
1E98	A	XRAY	1.9	0.19	0.24
1EBH	A	XRAY	1.9	0.19	1
1EEH	A	XRAY	1.9	0.23	0.27
1EFN	B	XRAY	2.5	0.21	0.28
1EFU	A	XRAY	2.5	0.17	0.28
1EFU	B	XRAY	2.5	0.17	0.28
1EHY	A	XRAY	2.1	0.19	0.23
1EWF	A	XRAY	1.7	0.2	0.25
1F13	A	XRAY	2.1	0.18	0.24
1FEH	A	XRAY	1.8	0.18	0.23

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
1FGK	A	XRAY	2	0.21	0.26
1FIN	B	XRAY	2.3	0.21	1
1FIP	A	XRAY	1.9	0.2	1
1FJM	A	XRAY	2.1	0.18	1
1FKD	A	XRAY	1.72	0.18	1
1FLE	I	XRAY	1.9	0.2	1
1FMT	A	XRAY	2	0.21	0.26
1FRO	A	XRAY	2.2	0.21	0.23
1G2A	A	XRAY	1.75	0.19	0.25
1GAR	A	XRAY	1.96	0.17	0.29
1GJM	A	XRAY	2.2	0.18	0.22
1GOT	A	XRAY	2	0.21	0.29
1GOT	G	XRAY	2	0.21	0.29
1GOT	B	XRAY	2	0.21	0.29
1GUA	B	XRAY	2	0.22	1
1GVP	A	XRAY	1.6	0.21	0.29
1HF8	A	XRAY	2	0.19	0.22
1HIA	I	XRAY	2.4	0.2	0.31
1HJR	A	XRAY	2.5	0.16	1
1HWG	A	XRAY	2.5	0.2	0.29
1HWG	B	XRAY	2.5	0.2	0.29
1HXP	A	XRAY	1.8	0.19	1
1ICW	A	XRAY	2.01	0.19	0.27
1ILR	1	XRAY	2.1	0.2	1
1IMB	A	XRAY	2.2	0.17	1
1ISA	A	XRAY	1.8	0.19	1
1IVY	A	XRAY	2.2	0.21	0.27
1JHG	A	XRAY	1.3	0.13	0.17
1JSG	A	XRAY	2.5	0.19	0.26
1KBA	A	XRAY	2.3	0.2	1
1KPF	A	XRAY	1.5	0.21	0.24
1KPT	A	XRAY	1.75	0.17	0.22
1KWA	A	XRAY	1.93	0.25	0.3
1M6P	A	XRAY	1.8	0.22	0.28
1MCT	A	XRAY	1.6	0.17	1
1MKB	A	XRAY	2	0.18	0.24
1MOR	A	XRAY	1.9	0.19	1
1MPG	A	XRAY	1.8	0.19	0.25
1NAW	A	XRAY	2	0.2	0.27
1NMB	N	XRAY	2.2	0.21	1
1NO3	A	XRAY	2.15	0.19	0.23
1NOX	A	XRAY	1.59	0.19	0.2

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
1NP4	A	XRAY	1.5	0.2	0.26
1NSE	A	XRAY	1.9	0.21	0.28
1NSY	A	XRAY	2	0.17	0.23
1OAC	A	XRAY	2	0.16	1
1OPY	A	XRAY	1.9	0.2	0.27
1OSP	O	XRAY	1.95	0.23	0.29
1PBG	A	XRAY	2.3	0.16	0.24
1PDA	A	XRAY	1.76	0.19	1
1PGT	A	XRAY	1.8	0.18	1
1QAZ	A	XRAY	1.78	0.18	0.23
1QCI	A	XRAY	2	0.23	1
1QFH	A	XRAY	2.2	0.22	0.27
1QHA	A	XRAY	2.25	0.21	0.28
1QHI	A	XRAY	1.9	0.23	0.29
1QJP	A	XRAY	1.65	0.15	0.2
1QME	A	XRAY	2.4	0.2	0.23
1QPA	A	XRAY	1.8	0.16	1
1QR2	A	XRAY	2.1	0.22	0.28
1QTQ	A	XRAY	2.25	0.24	0.25
1RBP	A	XRAY	2	0.18	1
1REG	X	XRAY	1.9	0.18	0.21
1RHS	A	XRAY	1.36	0.17	0.23
1RNE	A	XRAY	2.4	0.18	1
1RPO	A	XRAY	1.4	0.19	1
1SES	A	XRAY	2.5	0.18	1
1SHK	A	XRAY	1.9	0.17	0.22
1SLT	A	XRAY	1.9	0.17	1
1SMN	A	XRAY	2.04	0.17	1
1SMT	A	XRAY	2.2	0.22	0.25
1SOX	A	XRAY	1.9	0.17	0.22
1STF	I	XRAY	2.37	0.19	1
1TC1	A	XRAY	1.41	0.19	0.23
1THT	A	XRAY	2.1	0.23	1
1TOA	A	XRAY	1.8	0.18	0.2
1TOX	A	XRAY	2.3	0.23	0.31
1TRK	A	XRAY	2	0.16	1
1TX4	A	XRAY	1.65	0.17	0.21
1TX4	B	XRAY	1.65	0.17	0.21
1UBY	A	XRAY	2.4	0.2	1
1URP	A	XRAY	2.3	0.23	0.27
1UTG	A	XRAY	1.34	0.23	1
1VBT	A	XRAY	2.3	0.2	0.25

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
1VLB	A	XRAY	1.28	0.15	0.19
1VOK	A	XRAY	2.1	0.2	1
1XGS	A	XRAY	1.75	0.19	0.23
1XSO	A	XRAY	1.49	0.1	0.17
1YCS	A	XRAY	2.2	0.2	0.29
1YCS	B	XRAY	2.2	0.2	0.29
1YDR	E	XRAY	2.2	0.19	1
1YQV	L	XRAY	1.7	0.2	0.23
1YQV	H	XRAY	1.7	0.2	0.23
256B	A	XRAY	1.4	0.16	1
256L	A	XRAY	1.8	0.16	1
2ACY	A	XRAY	1.8	0.17	0.23
2ARC	A	XRAY	1.5	0.18	0.23
2ATJ	A	XRAY	2	0.18	0.2
2BC2	A	XRAY	1.7	0.2	0.25
2BLS	A	XRAY	2	0.22	1
2G3P	A	XRAY	1.9	0.26	0.3
2HDH	A	XRAY	2.2	0.2	0.25
2IHL	A	XRAY	1.4	0.17	1
2ILK	A	XRAY	1.6	0.16	1
2JEL	P	XRAY	2.5	0.21	0.28
2LIG	A	XRAY	2	0.18	1
2LIV	A	XRAY	2.4	0.18	1
2MBR	A	XRAY	1.8	0.2	0.26
2NAC	A	XRAY	1.8	0.15	1
2OHX	A	XRAY	1.8	0.17	1
2PCC	A	XRAY	2.3	0.17	1
2RN2	A	XRAY	1.48	0.2	1
2SCP	A	XRAY	2	0.18	1
2SHP	A	XRAY	2	0.2	0.27
2SIC	I	XRAY	1.8	0.18	1
2SPC	A	XRAY	1.8	0.2	1
2SQC	A	XRAY	2	0.15	0.19
2TCT	A	XRAY	2.1	0.18	1
2TGI	A	XRAY	1.8	0.17	1
2TPS	A	XRAY	1.25	0.18	0.22
2TRC	P	XRAY	2.4	0.19	0.28
2UGI	A	XRAY	2.2	0.23	0.28
3CLA	A	XRAY	1.75	0.16	1
3DAP	A	XRAY	2.2	0.17	0.23
3GBP	A	XRAY	2.4	0.16	1
3GRS	A	XRAY	1.54	0.19	1

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
3PFK	A	XRAY	2.4	0.17	1
3PMG	A	XRAY	2.4	0.16	0.19
3RN3	A	XRAY	1.45	0.22	1
3SDH	A	XRAY	1.4	0.16	1
3SGB	E	XRAY	1.8	0.12	1
3SGB	I	XRAY	1.8	0.12	1
4DFR	A	XRAY	1.7	0.15	1
4HTC	I	XRAY	2.3	0.17	1
5ACN	A	XRAY	2.1	0.21	1
5CPA	A	XRAY	1.54	0.19	1
5CPV	A	XRAY	1.6	0.19	1
5CSM	A	XRAY	2	0.19	0.24
5RUB	A	XRAY	1.7	0.18	1
6LDH	A	XRAY	2	0.2	1
6XIA	A	XRAY	1.65	0.14	1
7CAT	A	XRAY	2.5	0.21	1
830C	A	XRAY	1.6	0.21	0.27
8ATC	A	XRAY	2.5	0.17	1
8ATC	B	XRAY	2.5	0.17	1
8PRK	A	XRAY	1.85	0.19	0.23
8PTI	A	XRAY	1.8	0.16	1
9PAP	A	XRAY	1.65	0.16	1
9WGA	A	XRAY	1.8	0.17	1

There are 12 proteins with a total of 26 chains that have multiple chain entries in the 268 list. These 12 proteins were specifically characterized by Chakrabarti and Janin as heterodimers (1BRSA, 1BRSD, 1CSEE, 1CSEI, 1DANL, 1DANT, 1DANU, 1DHKA, 1DHKB, 1EFUA, 1EFUB, 1GOTA, 1GOTB, 1GOTG, 1HWGA, 1HWGB, 1TX4A, 1TX4B, 1YCSA, 1YCSB, 1YQVL, 1YQVH and 3SGBE, 3SGBI).[33] One heterodimeric chain 1YQVB does not meet the sequence identity requirement for PISCES applied to the original 70 heterodimeric protein set. An additional nonredundant heterodimeric protein 8ATC(A,B) is included from the 130 protein list.[26] Table 3.2 provides the protein length and the number of alignments for each of the 268 learning set list of proteins.

Table 3.2: Query length and number of alignments of individual proteins for the learning set list of 268 proteins.

PDB ID + Chain	No. of Query Residues	No. of Align.	PDB ID + Chain	No. of Query Residues	No. of Align.	PDB ID + Chain	No. of Query Residues	No. of Align.
12ASA	330	227	1CZJA	111	85	1RBPA	182	289
13PKA	415	1012	1DAAA	282	1000	1REGX	122	28
1A1IA	90	4981	1DANL	152	5215	1RHSA	296	1008
1A2KA	127	401	1DANT	80	109	1RNEA	340	1069
1A32A	88	982	1DANU	121	83	1RPOA	65	48
1A48A	306	1001	1DCSA	311	636	1SESA	421	1000
1A4IA	301	1001	1DFJI	457	4418	1SHKA	173	1000
1A4UA	254	1003	1DHKA	496	1021	1SLTA	134	733
1A6QA	382	1064	1DHKB	223	769	1SMNA	245	413
1AA7A	158	1000	1DHSA	361	420	1SMTA	122	1000
1ADDA	349	880	1DHTA	327	1001	1SOXA	466	1119
1ADEA	431	1000	1DINA	236	986	1STFI	98	132
1AF3A	196	513	1DMRA	823	1004	1TC1A	220	1000
1AFWA	393	1001	1DORA	311	1000	1THTA	305	85
1AG9A	175	747	1DPGA	485	1000	1TOAA	313	1020
1AH7A	245	271	1DQSA	393	1001	1TOXA	535	37
1AJSA	412	1000	1DYSA	348	205	1TRKA	680	1001
1AK0A	270	213	1E1KA	460	1008	1TX4A	198	1003
1AK4C	145	1000	1E5MA	416	1000	1TX4B	177	1001
1AKOA	268	1000	1E98A	215	1000	1UBYA	367	1001
1AL8A	359	1008	1EBHA	436	1000	1URPA	271	1000
1AMKA	251	1000	1EEHA	437	1008	1UTGA	70	55
1AMPA	291	1001	1EFNB	152	1000	1VBTA	165	1000
1AMUA	563	2309	1EFUA	385	1000	1VLBA	907	1188
1AN9A	340	485	1EFUB	282	1159	1VOKA	200	573
1AOBA	265	1007	1EHYA	294	1005	1XGSA	295	1000
1AORA	605	319	1EWFA	456	335	1XSOA	150	1003
1AQ0A	306	866	1F13A	731	419	1YCSA	199	418
1AQ6A	253	1000	1FEHA	574	1015	1YCSB	239	10257
1ATLA	202	1005	1FGKA	310	1003	1YDRE	350	1009
1AUOA	218	915	1FINB	260	1001	1YQVH	215	1092
1AVWB	177	342	1FIPA	98	1000	1YQVL	211	1003
1AW5A	340	1000	1FJMA	330	1002	256BA	106	58
1AW7A	194	86	1FKDA	107	1121	256LA	164	526
1AW9A	216	1000	1FLEI	57	272	2ACYA	98	796
1AYLA	541	626	1FMTA	314	1000	2ARCA	164	130
1AYXA	492	161	1FROA	183	1026	2ATJA	308	1002

PDB ID + Chain	No. of Query Residues	No. of Align.	PDB ID + Chain	No. of Query Residues	No. of Align.	PDB ID + Chain	No. of Query Residues	No. of Align.
1AZIA	153	1004	1G2AA	168	1000	2BC2A	227	1000
1B3AA	67	607	1GARA	212	1000	2BLSA	358	1000
1B5EA	246	563	1GJMA	414	1000	2G3PA	225	219
1B67A	68	312	1GOTA	350	1020	2HDHA	293	1195
1B8AA	438	1122	1GOTB	340	3674	2IHLA	129	916
1B8JA	449	896	1GOTG	73	222	2ILKA	160	240
1BA3A	550	1004	1GUAB	81	169	2JELP	85	1017
1BAMA	213	15	1GVPA	87	45	2LIGA	164	222
1BBHA	131	162	1HF8A	289	381	2LIVA	344	1004
1BD0A	388	1000	1HIAI	48	13	2MBRA	340	970
1BEAA	127	340	1HJRA	158	692	2NACA	393	1001
1BF2A	750	1029	1HWGA	191	1006	2OHXA	374	1005
1BFDA	528	1000	1HWGB	237	515	2PCCA	296	1515
1BG0A	356	1027	1HXPA	348	467	2RN2A	155	1000
1BIAA	321	1000	1ICWA	72	344	2SCPA	174	447
1BINA	143	415	1ILR1	152	247	2SHPA	525	1577
1BIQA	375	1001	1IMBA	277	1001	2SICI	107	55
1BISA	166	1000	1ISAA	192	1000	2SPCA	107	2661
1BJWA	382	1000	1IVYA	452	1099	2SQCA	631	970
1BLZA	331	1009	1JHGA	101	123	2TCTA	207	1002
1BMDA	327	1002	1JSGA	114	72	2TGIA	112	1000
1BN6A	294	1002	1KBAA	66	256	2TPSA	227	1002
1BO6A	297	990	1KPFA	126	1001	2TRCP	217	538
1BRSA	110	106	1KPTA	105	18	2UGIA	84	3
1BRSD	89	118	1KWAA	88	1340	3CLAA	213	303
1BRWA	433	849	1M6PA	152	112	3DAPA	320	137
1BSLA	324	1001	1MCTA	223	1030	3GBPA	307	1000
1BT3A	345	1046	1MKBA	171	951	3GRSA	478	1001
1BULA	265	1000	1MORA	368	1000	3PFKA	319	1222
1BUOA	121	1000	1MPGA	282	795	3PMGA	561	1004
1BXGA	356	1000	1NAWA	419	1005	3RN3A	124	626
1BXKA	355	1000	1NMBN	470	1000	3SDHA	146	633
1BXQA	323	1004	1NO3A	857	673	3SGBE	185	349
1BYOA	99	369	1NOXA	205	1000	3SGBI	56	572
1C02A	166	134	1NP4A	184	52	4DFRA	159	1000
1CB0A	283	989	1NSEA	444	371	4HTCI	65	35
1CDCA	99	186	1NSYA	271	1000	5ACNA	754	1009
1CEXA	214	272	1OACA	727	481	5CPAA	307	1075
1CG2A	393	1000	1OPYA	131	183	5CPVA	109	1232
1CHMA	401	1000	1OSPO	257	390	5CSMA	256	87

PDB ID + Chain	No. of Query Residues	No. of Align.	PDB ID + Chain	No. of Query Residues	No. of Align.	PDB ID + Chain	No. of Query Residues	No. of Align.
1CJXA	357	615	1PBGA	468	1002	5RUBA	490	1034
1CKIA	317	1004	1PDAA	313	1005	6LDHA	330	1001
1CMBA	104	82	1PGTA	210	1003	6XIAA	387	410
1CNZA	363	1000	1QAZA	351	60	7CATA	506	1002
1COZA	129	1096	1QCIA	262	436	830CA	168	1099
1CQXA	403	1003	1QFHA	212	3209	8ATCA	310	1000
1CRCA	105	1012	1QHAA	917	1408	8ATCB	153	226
1CRMA	260	1009	1QHIA	366	225	8PRKA	287	1001
1CRZA	403	2027	1QJPA	171	455	8PTIA	58	1543
1CSEE	274	1049	1QMEA	702	1001	9PAPA	212	1001
1CSEI	71	186	1QPAA	345	463	9WGAA	171	1556
1CSHA	435	1002	1QR2A	230	1000			
1CTTA	294	893	1QTQA	553	1003			

The frequency distribution plots for the characterization of the learning set list is shown in Figure 3.1. Here, 75891 query residues from 268 query protein chains were aligned to 235138 subject proteins for the generation of four histograms. Due to the absence of information related to entropy and density calculations, 2159 residues were eliminated before obtaining these graphs.

The frequency distribution plot of the query proteins to the protein length (Figure 3.1A) is a left-truncated normal distribution. The mean, median and mode of this distribution was found to be 283.18, 266.5 and 340, respectively. The two maxima correspond to the lengths 200 and 350 at which the frequency of proteins is 38 and 39 respectively. The length of query proteins range from 50 to 950 but, 95.90% of the proteins have length between 100 and 650. This means that the proteins of the 268 protein list are of average length, neither too short nor too long. The frequency distribution plot of 73732 query residues with respect to packing density (Figure 3.1B) can essentially be described as a normal distribution. The mean, median and mode of this distribution was found to be 14.43, 14 and 15, respectively. Packing density positions below 4 and above 25 have negligible residues, 0.054% and 0.26%

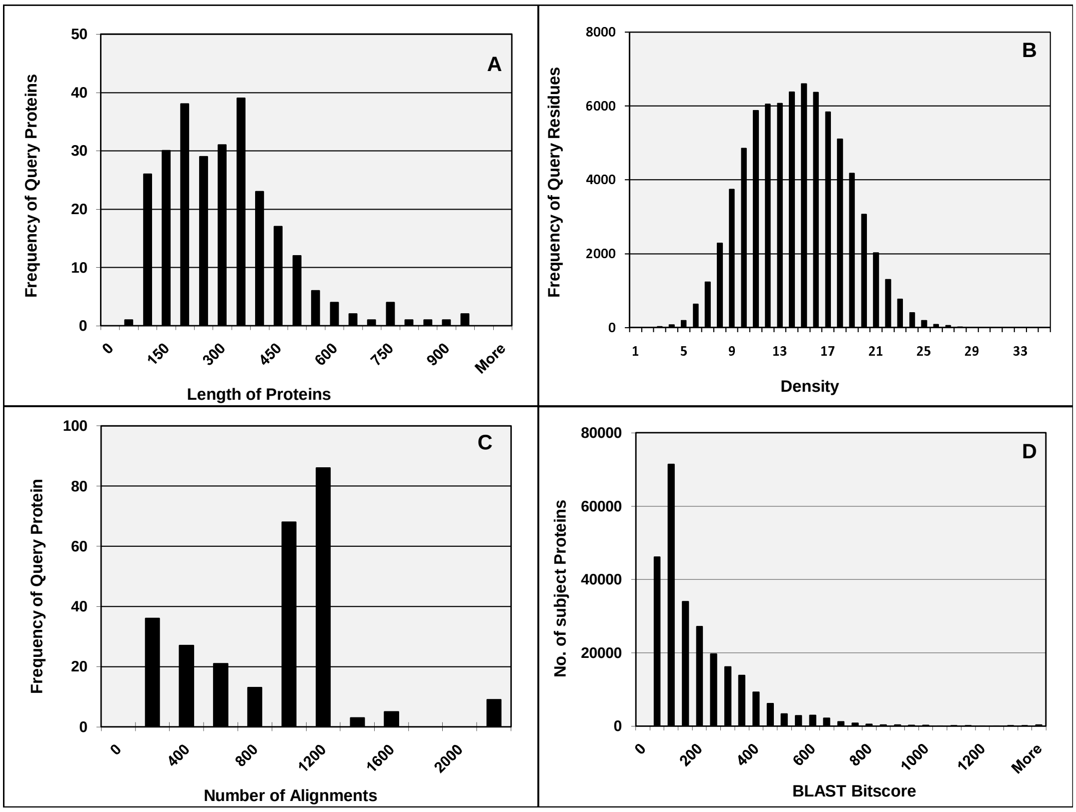


Figure 3.1: Frequency distributions for the characterization of learning set list of proteins. A total of 268 protein chains with 75891 query residues and a total of 235138 aligned subject protein sequences were used for these calculations. A. Frequency of query residues (ordinate) with respect to the length of each protein (abscissa) of the 268 learning set list. B. Frequency of 73732 query residues (ordinate) with respect to each packing density (abscissa). C. Frequency of query proteins (ordinate) is plotted against number of alignments (abscissa) obtained from NCBI BLASTP outputs for the learning set list. D. Frequency of 235138 aligned subject sequences (ordinate) with respect to BLAST bit scores (abscissa).

respectively, associated with them. Majority of the residues (58.75%) fall between packing density 11 and 17. The curve attains its maxima at packing density 15 where a total of 6602 residues show their presence. The number of alignments associated with the query proteins of the learning set list (Figure 3.1C) ranges from 3 for 2UGIA to $10,257$ for 1YCSB. Most of the residues (57.46%) have alignment lengths between $1000 - 1200$ and clustering at the extremes is not seen here. A right skewed histogram is obtained when the frequency of subject proteins is plotted against BLASTP bitscores for the learningset list (Figure 3.1D). A right skewed distribution is consistent with that for the randomized set of bitscores.[46] Bitscores range from $50 - 1400$ and have a total of 71372 subject proteins (27.46%) falling between bitscore value $50 - 100$.

3.1.2 Characterization of Monomeric List

The monomeric list has been characterized on the basis of the chain length of the respective proteins, their experimental determination method, resolution, R-factor, free R value, query length and number of alignments generated at each query residue position. The various frequency distribution plots of the query residues have also been noted here. Table 3.3 enlists the 75 monomeric proteins and the parameters that were used to cull the list in PISCES.

Table 3.3: Monomeric List of 75 proteins. PISCES culled list of 75 monomeric proteins. Each PDB ID obtained from chack 05 monomeric list of proteins, was culled by considering individual protein chains. The percentage sequence identity cutoff of $\leq 25\%$, a resolution of $0.0 - 2.5\AA$, R-factor ≤ 0.3 and sequence length $40 - 10000$ was used for culling.

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
13PK	A	XRAY	2.5	0.22	0.29
1A7V	A	XRAY	2.3	0.19	0.24
1AFK	A	XRAY	1.7	0.21	0.27
1AG9	A	XRAY	1.8	0.2	0.25
1AH7	A	XRAY	1.5	0.2	0.23
1AKO	A	XRAY	1.7	0.17	0.2

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
1AMU	A	XRAY	1.9	0.21	0.25
1ATL	A	XRAY	1.8	0.16	1
1AW7	A	XRAY	1.95	0.18	1
1AYL	A	XRAY	1.8	0.2	0.23
1BEA	A	XRAY	1.95	0.2	0.29
1BIN	A	XRAY	2.2	0.2	0.3
1BKZ	A	XRAY	1.9	0.19	0.26
1BYO	A	XRAY	2	0.19	0.23
1C02	A	XRAY	1.8	0.2	0.25
1CKI	A	XRAY	2.3	0.19	0.28
1CLU	A	XRAY	1.7	0.2	0.26
1CQX	A	XRAY	1.75	0.18	0.21
1DYS	A	XRAY	1.6	0.18	0.24
1E0S	A	XRAY	2.28	0.17	0.23
1EHY	A	XRAY	2.1	0.19	0.23
1EPA	A	XRAY	2.1	0.2	1
1EWF	A	XRAY	1.7	0.2	0.25
1FEH	A	XRAY	1.8	0.18	0.23
1FGK	A	XRAY	2	0.21	0.26
1FJM	A	XRAY	2.1	0.18	1
1FKD	A	XRAY	1.72	0.18	1
1FMT	A	XRAY	2	0.21	0.26
1G2A	A	XRAY	1.75	0.19	0.25
1GAR	A	XRAY	1.96	0.17	0.29
1GJM	A	XRAY	2.2	0.18	0.22
1HF8	A	XRAY	2	0.19	0.22
1ILR	1	XRAY	2.1	0.2	1
1KPT	A	XRAY	1.75	0.17	0.22
1KWA	A	XRAY	1.93	0.25	0.3
1MPG	A	XRAY	1.8	0.19	0.25
1MSS	A	XRAY	2.4	0.2	1
1NAW	A	XRAY	2	0.2	0.27
1NP4	A	XRAY	1.5	0.2	0.26
1PBG	A	XRAY	2.3	0.16	0.24
1PDA	A	XRAY	1.76	0.19	1
1PPO	A	XRAY	1.8	0.15	1
1QAZ	A	XRAY	1.78	0.18	0.23
1QCI	A	XRAY	2	0.23	1
1QDM	A	XRAY	2.3	0.22	1
1QHA	A	XRAY	2.25	0.21	0.28
1QJP	A	XRAY	1.65	0.15	0.2
1QME	A	XRAY	2.4	0.2	0.23

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
1QPA	A	XRAY	1.8	0.16	1
1QTQ	A	XRAY	2.25	0.24	0.25
1RB3	A	XRAY	2.3	0.16	1
1RHS	A	XRAY	1.36	0.17	0.23
1SHK	A	XRAY	1.9	0.17	0.22
1THT	A	XRAY	2.1	0.23	1
1TOA	A	XRAY	1.8	0.18	0.2
1TON	A	XRAY	1.8	0.2	1
1URP	A	XRAY	2.3	0.23	0.27
1VBT	A	XRAY	2.3	0.2	0.25
1XGS	A	XRAY	1.75	0.19	0.23
256B	A	XRAY	1.4	0.16	1
256L	A	XRAY	1.8	0.16	1
2ACY	A	XRAY	1.8	0.17	0.23
2ATJ	A	XRAY	2	0.18	0.2
2BC2	A	XRAY	1.7	0.2	0.25
2BLS	A	XRAY	2	0.22	1
2G3P	A	XRAY	1.9	0.26	0.3
2IHL	A	XRAY	1.4	0.17	1
2MBR	A	XRAY	1.8	0.2	0.26
2SCP	A	XRAY	2	0.18	1
2SHP	A	XRAY	2	0.2	0.27
2TPS	A	XRAY	1.25	0.18	0.22
2UGI	A	XRAY	2.2	0.23	0.28
3PMG	A	XRAY	2.4	0.16	0.19
830C	A	XRAY	1.6	0.21	0.27
8PTI	A	XRAY	1.8	0.16	1

Note that only 65 monomeric protein chains (see Appendix Table C.1) are just the subset of the 268 protein list. The monomeric list of 75 protein chains were obtained as a result of applying PISCES to the original set of 103 homodimeric proteins.[34] The protein length and the number of alignments have been compiled in Table 3.4 for the monomeric list of 75 proteins.

Table 3.4: Query length and number of alignments of individual proteins for the list of 75 monomeric proteins.

PDB ID + Chain	No. of Query Residues	No. of Align.	PDB ID + Chain	No. of Query Residues	No. of Align.	PDB ID + Chain	No. of Query Residues	No. of Align.
13PKA	415	1012	1FJMA	330	1002	1RB3A	159	1000
1A7VA	125	197	1FKDA	107	1121	1RHSA	296	1008
1AFKA	124	615	1FMTA	314	1000	1SHKA	173	1000
1AG9A	175	747	1G2AA	168	1000	1THTA	305	85
1AH7A	245	271	1GARA	212	1000	1TOAA	313	1020
1AKOA	268	1000	1GJMA	414	1000	1TONA	235	1012
1AMUA	563	2309	1HF8A	289	381	1URPA	271	1000
1ATLA	202	1005	1ILR1	152	247	1VBTA	165	1000
1AW7A	194	86	1KPTA	105	18	1XGSA	295	1000
1AYLA	541	619	1KWAA	88	1340	256BA	106	58
1BEAA	127	340	1MPGA	282	795	256LA	164	526
1BINA	143	415	1MSSA	243	1000	2ACYA	98	796
1BKZA	135	920	1NAWA	419	1005	2ATJA	308	1002
1BYOA	99	369	1NP4A	184	52	2BC2A	227	1000
1C02A	166	134	1PBGA	468	1002	2BLSA	358	1000
1CKIA	317	1004	1PDAA	313	1005	2G3PA	225	219
1CLUA	166	1001	1PPOA	216	1001	2IHLA	129	916
1CQXA	403	1003	1QAZA	351	60	2MBRA	340	970
1DYSA	348	205	1QCIA	262	436	2SCPA	174	447
1E0SA	174	1001	1QDMA	478	1763	2SHPA	525	1577
1EHYA	294	1005	1QHAA	917	1408	2TPSA	227	1002
1EPAA	164	295	1QJPA	171	455	2UGIA	84	3
1EWFA	456	335	1QMEA	702	1001	3PMGA	561	1004
1FEHA	574	1015	1QPAA	345	463	830CA	168	1099
1FGKA	310	1003	1QTQA	553	1003	8PTIA	58	1543

The frequency distribution plots of the monomeric list for various parameters are shown in Figure 3.2. The frequency distributions have been obtained from a total of 20076 query residues belonging to 75 monomeric proteins. The total number of alignments generated for this set was 59751.

The residue frequency and protein length plot of query proteins show a mixed character (Figure 3.2A). The length of proteins range from $100 - 950$. Only a few proteins (2.67%) have length greater than 600 residues and approximately 21.33%

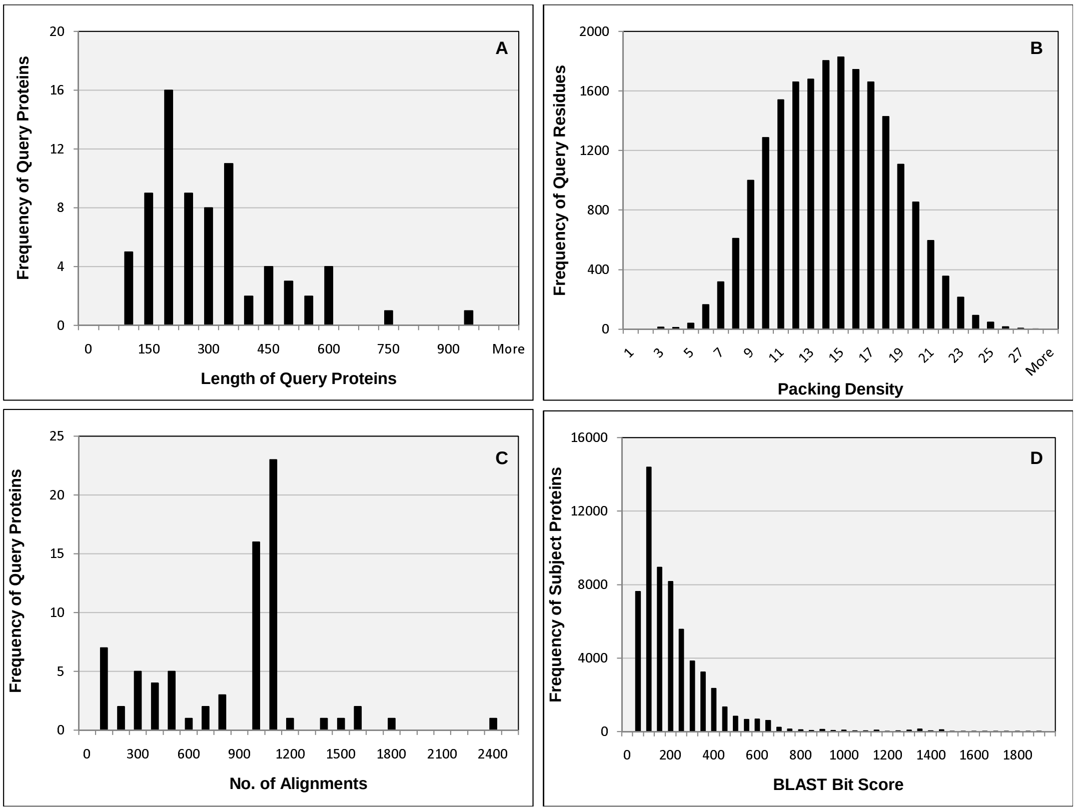


Figure 3.2: Frequency distribution plots for the characterisation of monomeric list of proteins. A total of 75 monomeric proteins with 20076 query residues and a total of 59751 aligned subject protein sequences were used for these calculations. A. Frequency of query protein (ordinate) have been plotted against their length (abscissa). B. Number of query residues (ordinate) at each packing density position (abscissa) has been plotted here. C. Frequency of query proteins (ordinate) versus the number of BLASTP alignments (abscissa) and D. Frequency of subject proteins (ordinate) with respect to BLAST bit scores (abscissa).

of query proteins have the maximum frequency at protein length 200. The density distribution plot of the monomeric query residues (Figure 3.2B) show a normal distribution that ranges from density $3 - 28$. The curve attains its maxima at density 14 and 15 at which 18.08% residues are found. The histogram for the number of alignments of the monomeric list proteins has been shown in (Figure 3.2C). It ranges from $100 - 2400$ alignments with maximum frequency of 52% between $1000 - 1200$. The histogram plotted for the frequency of subject proteins versus BLAST bit score is a right skewed normal distribution (Figure 3.2D). It ranges from $50 - 1900$ and attains its maxima at bit score value equal to 100 where its frequency is 14382 (24.07% of the total residues).

3.1.3 Characterization of Homodimeric List

The homodimeric list consists of 106 proteins that were aligned to a total of 83048 subject proteins. From a total of 29427 query residues, a total of 28734 residues were processed because the remaining 693 residues were eliminated due to non-availability of information from either the mmCIF files or the BLASTP output files. The homodimeric list of 106 proteins was obtained as a result of applying PISCES to the original set of 122 homodimeric proteins that were specifically characterized by Bahadur et al., 2003.[34] Here, the complete list is characterized according to the standard set of parameters (discussed in Chapter 2), the length of the constituent proteins and their number of alignments. The frequency distribution plots for various parameters have also been presented here.

Table 3.5 summarizes the homodimeric list, their protein database (RCSB) identifiers, chain identifiers, mode of structure determination, resolution, R-factor and free R value.

Table 3.5: Homodimeric list of 106 proteins. Chain culled list of 106 homodimeric proteins. Each protein chain shares a percentage sequence identity of $\leq 25\%$, has a resolution between $0.0 - 2.5\AA$, R-factor ≤ 0.3, and sequence length between $40 - 10000$.

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
12AS	A	XRAY	2.2	0.16	0.29
1A4I	A	XRAY	1.5	0.2	0.23
1A4U	A	XRAY	1.92	0.2	0.24
1AA7	A	XRAY	2.08	0.21	0.28
1ADE	A	XRAY	2	0.2	1
1AFW	A	XRAY	1.8	0.19	0.24
1AJS	A	XRAY	1.6	0.17	1
1AMK	A	XRAY	1.83	0.11	1
1AOR	A	XRAY	2.3	0.15	1
1AQ6	A	XRAY	1.95	0.19	0.25
1AUO	A	XRAY	1.8	0.21	0.27
1B3A	A	XRAY	1.6	0.17	0.24
1B5E	A	XRAY	1.6	0.19	0.21
1B67	A	XRAY	1.48	0.19	0.27
1B8A	A	XRAY	1.9	0.17	0.2
1B8J	A	XRAY	1.9	0.18	0.2
1BAM	A	XRAY	1.95	0.19	1
1BBH	A	XRAY	1.8	0.18	1
1BD0	A	XRAY	1.6	0.24	0.27
1BIF	A	XRAY	2	0.18	0.25
1BIQ	A	XRAY	2.05	0.19	0.26
1BIS	A	XRAY	1.95	0.2	0.26
1BJW	A	XRAY	1.8	0.21	0.27
1BMD	A	XRAY	1.9	0.15	1
1BRW	A	XRAY	2.1	0.23	0.28
1BSL	A	XRAY	1.95	0.19	1
1BSR	A	XRAY	1.9	0.18	1
1BUO	A	XRAY	1.9	0.21	0.25
1BXG	A	XRAY	2.3	0.17	1
1BXK	A	XRAY	1.9	0.2	1
1CDC	A	XRAY	2	0.19	1
1CG2	A	XRAY	2.5	0.2	0.22
1CHM	A	XRAY	1.9	0.18	1
1CMB	A	XRAY	1.8	0.19	1
1CNZ	A	XRAY	1.76	0.2	0.26
1COZ	A	XRAY	2	0.2	0.26
1CSH	A	XRAY	1.65	0.16	1

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
1CTT	A	XRAY	2.2	0.19	1
1CVU	A	XRAY	2.4	0.2	0.23
1CZJ	A	XRAY	2.16	0.2	0.26
1DAA	A	XRAY	1.94	0.18	1
1DOR	A	XRAY	2	0.17	0.21
1DPG	A	XRAY	2	0.21	0.26
1DQS	A	XRAY	1.8	0.17	0.22
1E98	A	XRAY	1.9	0.19	0.24
1EBH	A	XRAY	1.9	0.19	1
1F13	A	XRAY	2.1	0.18	0.24
1FIP	A	XRAY	1.9	0.2	1
1FRO	A	XRAY	2.2	0.21	0.23
1GVP	A	XRAY	1.6	0.21	0.29
1HJR	A	XRAY	2.5	0.16	1
1HSS	A	XRAY	2.06	0.19	0.22
1HXP	A	XRAY	1.8	0.19	1
1ICW	A	XRAY	2.01	0.19	0.27
1IMB	A	XRAY	2.2	0.17	1
1ISA	A	XRAY	1.8	0.19	1
1IVY	A	XRAY	2.2	0.21	0.27
1JHG	A	XRAY	1.3	0.13	0.17
1JSG	A	XRAY	2.5	0.19	0.26
1KBA	A	XRAY	2.3	0.2	1
1KPF	A	XRAY	1.5	0.21	0.24
1M6P	A	XRAY	1.8	0.22	0.28
1MKB	A	XRAY	2	0.18	0.24
1MOR	A	XRAY	1.9	0.19	1
1NOX	A	XRAY	1.59	0.19	0.2
1NSE	A	XRAY	1.9	0.21	0.28
1NSY	A	XRAY	2	0.17	0.23
1OAC	A	XRAY	2	0.16	1
1OPY	A	XRAY	1.9	0.2	0.27
1PGT	A	XRAY	1.8	0.18	1
1QFH	A	XRAY	2.2	0.22	0.27
1QHI	A	XRAY	1.9	0.23	0.29
1QR2	A	XRAY	2.1	0.22	0.28
1REG	X	XRAY	1.9	0.18	0.21
1RPO	A	XRAY	1.4	0.19	1
1SES	A	XRAY	2.5	0.18	1
1SLT	A	XRAY	1.9	0.17	1
1SMN	A	XRAY	2.04	0.17	1
1SMT	A	XRAY	2.2	0.22	0.25

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
1SOX	A	XRAY	1.9	0.17	0.22
1TC1	A	XRAY	1.41	0.19	0.23
1TOX	A	XRAY	2.3	0.23	0.31
1TRK	A	XRAY	2	0.16	1
1UBY	A	XRAY	2.4	0.2	1
1UTG	A	XRAY	1.34	0.23	1
1VOK	A	XRAY	2.1	0.2	1
1WTL	A	XRAY	1.9	0.16	1
1XSO	A	XRAY	1.49	0.1	0.17
2ARC	A	XRAY	1.5	0.18	0.23
2HDH	A	XRAY	2.2	0.2	0.25
2ILK	A	XRAY	1.6	0.16	1
2LIG	A	XRAY	2	0.18	1
2NAC	A	XRAY	1.8	0.15	1
2OHX	A	XRAY	1.8	0.17	1
2SPC	A	XRAY	1.8	0.2	1
2SQC	A	XRAY	2	0.15	0.19
2TCT	A	XRAY	2.1	0.18	1
2TGI	A	XRAY	1.8	0.17	1
3DAP	A	XRAY	2.2	0.17	0.23
3GRS	A	XRAY	1.54	0.19	1
3SDH	A	XRAY	1.4	0.16	1
3SSI	A	XRAY	2.3	0.18	1
5CSM	A	XRAY	2	0.19	0.24
5RUB	A	XRAY	1.7	0.18	1
8PRK	A	XRAY	1.85	0.19	0.23
9WGA	A	XRAY	1.8	0.17	1

Note that only 99 homodimeric protein chains (see Appendix Table C.2) are just the subset of the 268 protein list. The homodimeric list of 106 protein chains were obtained as a result of applying PISCES to the original set of 122 homodimeric proteins.[34] Table 3.6 enlists the protein length and their number of alignments along with their combined protein database (RCSB) identifier and chain identifiers for the homodimeric list of proteins.

Table 3.6: Query length and number of alignments of individual proteins for the list of 106 homodimeric proteins

PDB ID + Chain	No. of Query Residues	No. of Align.	PDB ID + Chain	No. of Query Residues	No. of Align.	PDB ID + Chain	No. of Query Residues	No. of Align.
12ASA	330	227	1COZA	129	1096	1QFHA	212	3209
1A4IA	301	1001	1CSHA	435	1002	1QHIA	366	225
1A4UA	254	1003	1CTTA	294	893	1QR2A	230	1000
1AA7A	158	1000	1CVUA	552	1731	1REGX	122	28
1ADEA	431	1000	1CZJA	111	85	1RPOA	65	48
1AFWA	393	1001	1DAAA	282	1000	1SESA	421	1000
1AJSA	412	1000	1DORA	311	1000	1SLTA	134	733
1AMKA	251	1000	1DPGA	485	1000	1SMNA	245	413
1AORA	605	319	1DQSA	393	1001	1SMTA	122	1000
1AQ6A	253	1000	1E98A	215	1000	1SOXA	466	1119
1AUOA	218	915	1EBHA	436	1000	1TC1A	220	1000
1B3AA	67	607	1F13A	731	419	1TOXA	535	37
1B5EA	246	563	1FIPA	98	1000	1TRKA	680	1001
1B67A	68	312	1FROA	183	1026	1UBYA	367	1001
1B8AA	438	1122	1GVPA	87	45	1UTGA	70	55
1B8JA	449	896	1HJRA	158	692	1VOKA	200	573
1BAMA	213	15	1HSSA	124	248	1WTLA	108	1014
1BBHA	131	162	1HXPA	348	467	1XSOA	150	1003
1BD0A	388	1000	1ICWA	72	344	2ARCA	164	130
1BIFA	469	1069	1IMBA	277	1001	2HDHA	293	1195
1BIQA	375	1001	1ISAA	192	1000	2ILKA	160	240
1BISA	166	1000	1IVYA	452	1099	2LIGA	164	222
1BJWA	382	1000	1JHGA	101	123	2NACA	393	1001
1BMDA	327	1002	1JSGA	114	72	2OHXA	374	1005
1BRWA	433	849	1KBAA	66	256	2SPCA	107	2661
1BSLA	324	1001	1KPFA	126	1001	2SQCA	631	970
1BSRA	124	623	1M6PA	152	112	2TCTA	207	1002
1BUOA	121	1000	1MKBA	171	951	2TGIA	112	1000
1BXGA	356	1000	1MORA	368	1000	3DAPA	320	137
1BXKA	355	1000	1NOXA	205	1000	3GRSA	478	1001
1CDCA	99	186	1NSEA	444	371	3SDHA	146	633
1CG2A	393	1000	1NSYA	271	1000	3SSIA	113	66
1CHMA	401	1000	1OACA	727	481	5CSMA	256	87
1CMBA	104	82	1OPYA	131	183	5RUBA	490	1034
1CNZA	363	1000	1PGTA	210	1003	8PRKA	287	1001
9WGAA	171	1546						

The frequency distribution plots for the homodimeric list of proteins has been shown in Figure 3.3. The four histograms shown here were obtained from the information provided by 28734 aligned query residues belonging to 106 proteins. The total number of alignments generated for this set was 83048.

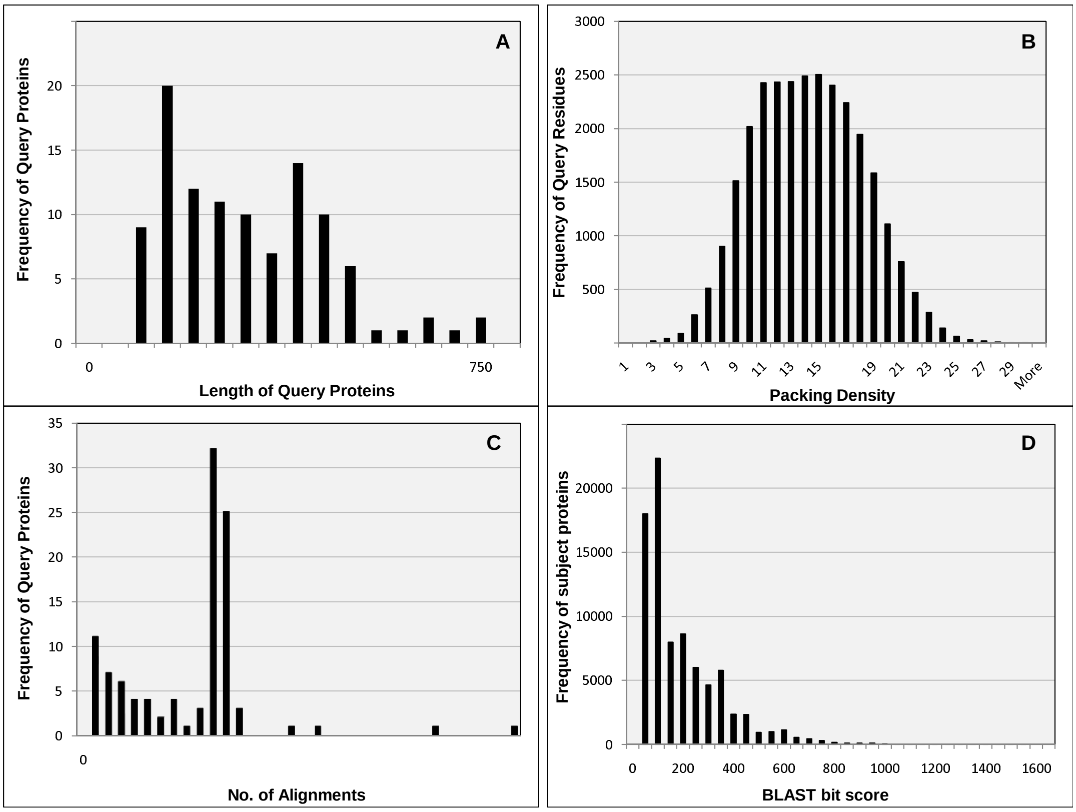


Figure 3.3: Frequency distribution plots for the characterization of homodimeric list of proteins. A total of 106 homodimeric proteins with 28734 aligned query residues and a total of 83048 aligned subject protein sequences were used for these plots. A. Frequency of query proteins (ordinate) have been plotted against their length (abscissa). B. Number of query residues (ordinate) at each packing density (abscissa) position has been plotted here. C. Frequency of query proteins (ordinate) versus the number of BLASTP alignments (abscissa) and D. Frequency of subject proteins (ordinate) with respect to BLAST bit scores (abscissa).

The histogram (Figure 3.3A) shows the number of query proteins as a function of their length. The plot is not a normal distribution and the length of the proteins of the homodimeric list range from $100 - 750$. The maximum number of proteins (18.87%)

have a length between $100-150$. Only a few proteins (6.604%) have residues more than 500. The packing density frequency distribution curve shown in Figure 3.3B is normal in character. It ranges from density $3-30$. Most of the residues (87.40%) lie between density $9-20$. The occupancy of residues beyond density 27 is minimal (0.05%). The frequency of query proteins versus number of alignments has been plotted in Figure 3.5C. The curve does not represent a normal distribution and the number of alignments range from $100-3300$. The maximum number of proteins (53.77%) have alignments lying between $1000-1100$. Only 4 proteins (3.77%) have number of alignments larger than 1200. The frequency of subject proteins versus BLAST bit score histogram is essentially a right skewed normal distribution shown in Figure 3.3D. It ranges from bit scores $50-1550$ and attains maxima between bit score value $50-100$. 26.92% of the subject proteins have their bit score values in that range. The bit score values beyond 950 are associated with minimal residues (0.17%). A right skewed distribution of blast bit scores was obtained which is consistent with the randomized set of bit score distribution.

3.1.4 Characterization of Heterodimeric List

In this section, the heterodimeric list obtained as discussed in Chapter 2 has been characterized on the basis of various parameters like protein database and chain identifiers, resolution, R-factor, free R value, number of alignments, protein length and frequency distribution plots for a set of parameters. The 50 protein chains belonging to this list along with their PISCES culling parameters have been enlisted in Table 3.7.

Table 3.7: Heterodimeric list of 50 protein chains. A list of 50 chain culled heterodimeric protein chains. All the proteins in the list have a resolution of $0.0-2.5\mathring{A}$ and a sequence length of $40-10000$. They share a sequence percentage identity of $\leq 25\%$ and have R-factor ≤ 0.3.

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
1A2K	A	XRAY	2.5	0.21	0.27

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
1AK4	A	XRAY	2.36	0.24	0.31
1AK4	C	XRAY	2.36	0.24	0.31
1AVW	B	XRAY	1.75	0.19	0.21
1BRS	A	XRAY	2	0.17	1
1BRS	D	XRAY	2	0.17	1
1CSE	E	XRAY	1.2	0.18	1
1CSE	I	XRAY	1.2	0.18	1
1DAN	U	XRAY	2	0.19	0.22
1DAN	T	XRAY	2	0.19	0.22
1DAN	L	XRAY	2	0.19	0.22
1DFJ	E	XRAY	2.5	0.19	1
1DFJ	I	XRAY	2.5	0.19	1
1DHK	B	XRAY	1.85	0.18	0.22
1DHK	A	XRAY	1.85	0.18	0.22
1EFN	B	XRAY	2.5	0.21	0.28
1EFU	B	XRAY	2.5	0.17	0.28
1EFU	A	XRAY	2.5	0.17	0.28
1FIN	B	XRAY	2.3	0.21	1
1FLE	I	XRAY	1.9	0.2	1
1GOT	G	XRAY	2	0.21	0.29
1GOT	B	XRAY	2	0.21	0.29
1GOT	A	XRAY	2	0.21	0.29
1GUA	B	XRAY	2	0.22	1
1HIA	I	XRAY	2.4	0.2	0.31
1HWG	B	XRAY	2.5	0.2	0.29
1HWG	A	XRAY	2.5	0.2	0.29
1KB5	B	XRAY	2.5	0.22	1
1MCT	A	XRAY	1.6	0.17	1
1NCA	L	XRAY	2.5	0.19	1
1NMB	N	XRAY	2.2	0.21	1
1OSP	O	XRAY	1.95	0.23	0.29
1STF	I	XRAY	2.37	0.19	1
1STF	E	XRAY	2.37	0.19	1
1TX4	A	XRAY	1.65	0.17	0.21
1TX4	B	XRAY	1.65	0.17	0.21
1VFB	B	XRAY	1.8	0.18	1
1VFB	C	XRAY	1.8	0.18	1
1YCS	A	XRAY	2.2	0.2	0.29
1YCS	B	XRAY	2.2	0.2	0.29
1YDR	E	XRAY	2.2	0.19	1
2PCC	A	XRAY	2.3	0.17	1
2PCC	B	XRAY	2.3	0.17	1

PDB ID	Chain	Exptl.	Resolution	R-factor	FreeRvalue
2PTC	I	XRAY	1.9	0.19	1
2SIC	I	XRAY	1.8	0.18	1
2TRC	P	XRAY	2.4	0.19	0.28
3SGB	E	XRAY	1.8	0.12	1
3SGB	I	XRAY	1.8	0.12	1
4CPA	A	XRAY	2.5	0.2	1
4HTC	I	XRAY	2.3	0.17	1

Note that only 40 heterodimeric protein chains (see Appendix Table C.3) are just the subset of the 268 protein list. The heterodimeric list of 50 protein chains were obtained as a result of applying PISCES to the original set of 70 heterodimeric proteins.[33] The query protein length and the number of alignments have been enlisted in Table 3.8 for each protein identifier belonging to the 50 protein chains of the heterodimeric list.

Table 3.8: Query length and number of alignments of individual proteins for the list of 50 heterodimeric protein chains.

PDB ID + Chain	No. of Query Residues	No. of Align.	PDB ID + Chain	No. of Query Residues	No. of Align.	PDB ID + Chain	No. of Query Residues	No. of Align.
1A2KA	127	401	1EFUB	282	1159	1TX4A	198	1003
1AK4A	165	1000	1FINB	260	1001	1TX4B	177	1001
1AK4C	145	1000	1FLEI	57	272	1VFBB	116	1001
1AVWB	177	342	1GOTA	350	1020	1VFBC	129	1071
1BRSA	110	106	1GOTB	340	3674	1YCSA	199	418
1BRSD	89	118	1GOTG	73	222	1YCSB	239	10257
1CSEE	274	1049	1GUAB	81	169	1YDRE	350	1009
1CSEI	71	186	1HIAI	48	13	2PCCA	296	1515
1DANL	152	5215	1HWGA	191	1006	2PCCB	108	1010
1DANT	80	109	1HWGB	237	515	2PTCI	58	1718
1DANU	121	83	1KB5B	117	1012	2SICI	107	55
1DFJE	124	628	1MCTA	223	1030	2TRCP	217	538
1DFJI	457	4418	1NCAL	214	1002	3SGBE	185	349
1DHKA	496	1021	1NMBN	470	1000	3SGBI	56	572
1DHKB	223	769	1OSPO	257	390	4CPAA	307	1068
1EFNB	152	1000	1STFE	212	1001	4HTCI	65	35
1EFUA	385	1000	1STFI	98	132			

The frequency distribution plots for the heterodimeric list of protein chains have been plotted in Figure 3.4. A total of 9665 residues belonging to the 50 heterodimeric protein chains that were aligned to a total of 55683 subject proteins have been processed for plotting these histograms. But only a total of 9230 aligned residues were used for obtaining these histograms. Due to the unavailability of coordinate information from the mmCIF files or the absence of residue positions in the BLASTP output files, or the presence of flagged residue positions in the entropy files, a total of 435 residues were eliminated.

The histogram obtained by plotting the number of proteins at each protein length increment, shown in Figure 3.4A, ranges from $50 - 500$. The graph is a right skewed normal distribution. Most of the proteins (40%) have length between $50 - 150$. The density distribution plot (Figure 3.4B) for the heterodimers is essentially a normal distribution that ranges from packing density increment $3 - 31$. Maximal occupancy of 9.48% and 8.99% are associated with packing density increments $14 - 16$. Residue occupancy below packing density 3 and above packing density 26 is nominal (0.15%). The frequency distribution representing number of query proteins as a function of the number of alignments is shown in Figure 3.4C. The number of alignments range from $500 - 10500$. Most of the proteins (92%) have number of aligned subject proteins below 2000. Two maxima, obtained at $0 - 500$ alignments and $1000 - 1500$ have a protein occupancy of 34% each. Only 8% of the proteins are aligned to more than 2000 subject proteins. The number of subject proteins at each BLAST bit score increment has been presented in Figure 3.4D. The graph looks like a left truncated normal distribution. It attains its maxima between BLAST bit scores 50 and 100 which is characterized by an occupancy of 16671 subject proteins (29.94%). The BLAST bit score values range from $50 - 1050$ and have a nominal occupancy of 0.49% above BLAST score value 700.

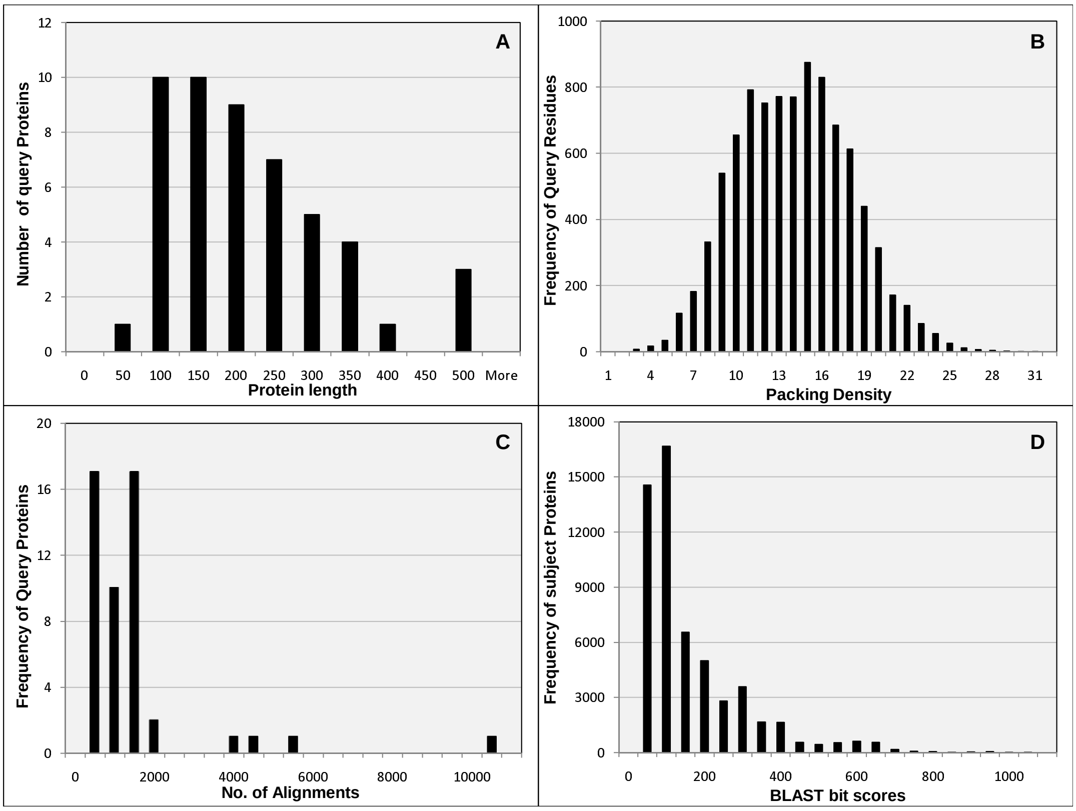


Figure 3.4: Frequency distribution plots for the characterization of heterodimeric list of protein chains. A total of 50 heterodimeric protein chains with 9665 query residues and a total of 55683 aligned subject protein sequences were used for these calculations. A. Frequency of query proteins (ordinate) have been plotted against their length (abscissa). B. Number of query residues (ordinate) at each packing density position (abscissa) has been plotted here. C. Frequency of query proteins (ordinate) versus the number of BLASTP alignments (abscissa) and D. Frequency of subject proteins (ordinate) with respect to BLAST bit scores (abscissa).

3.2 Aggregate Correlation Plots

In this section, various aggregate correlation plots have been summarized for the four protein lists. Overlay plot for the comparison of entropy values obtained from different methods, packing density versus relative surface accessibility plots and correlation plots for the various homology-based parameters have been evaluated here. Comparisons of homology based parameters of the proteins have been performed at two levels. In the first level, the homology-based parameters of proteins in the mixed aggregate plot (consisting of proteins in monomeric, homodimeric and heterodimeric proteins) have been compared. In the second level, comparisons of different protein trends in each of these lists are carried out individually. Residue loss after the processing of aggregate set of aligned residues, was found to be 2164 for the learning set list (see subsection 3.1.1), 700 for the monomeric list (see subsection 3.1.2), 690 for the homodimeric list (see subsection 3.1.3) and 435 for the heterodimeric list (see subsection 3.1.4). Hence, for the double average correlation plots and the frequency distribution plots, a total of 73727 residues of the learning set list, 20075 residues of the monomeric list, 28733 residues of the homodimeric list and 9230 residues of the heterodimeric list were processed.

3.2.1 Validation of Packing Density Calculation

The packing density calculation was verified by two independent methods. First by comparison of inverse packing density trend with residue hydrophobicity and residue hydrophilicity trends. Second, by comparing the aggregate density of the four protein lists with their aggregate relative surface accessibility values obtained from NACCESS[31] for the specified protein chains.

For investigating the density-hydrophobicity correlation, hydrophobicity scale was selected. Figure 3.5A represents the aggregate hydrophobicity trends of 268 protein list for various hydrophobicity scales, like, Hopp_Woods,[47] Engelman_Steitz[48] , Sharp_Honig[49] and Miyazawa_Jernigan,[50] as a function of inverse packing density. Two sharp peaks that compare closely were obtained from Sharp_Honig and Miyazawa_Jernigan scales. Due to its widespread use in the prediction of retention time in hydrophobic interaction chromatography, Miyazawa_Jernigan hydrophobicity scale was chosen for this work.

Figure 3.5B represents an overlay plot of some homology-based parameters like entropy, fraction hydrophilic, fraction hydrophobic, fraction non-hydrophobic and fraction gaps. The trends of fraction hydrophobic and fraction non-hydrophobic were found to be the inverse of each other. The overall aggregate trend of entropy and fraction hydrophilic were found to be comparable in terms of their relation with inverse packing density. The aggregate curve obtained from the fraction gaps analysis shows a linear increase with increase in inverse packing density and hence appears to be less helpful in providing any insights towards justifying packing density calculations. Hence, fraction hydrophobic and fraction hydrophilic were chosen for this study.

Figure 3.6A represents an aggregate overlay plot of entropy (plotted on primary axis) and fraction hydrophilic (plotted on secondary axis) for the four protein lists, 268 list, monomeric, homodimeric and heterodimeric lists. The hydrophilic fraction is expected to be low in the core of the protein where the packing density is high and the hydrophilic fraction is expected to be higher as we proceed towards the surface (where the packing density is low) from the protein's core. As is expected, the hydrophilic fraction increases with increase in inverse packing density. Figure 3.6B represents the aggregate overlay plot of entropy (primary axis) and fraction hydrophobic (secondary axis) for the learning set list (268 protein chains), monomeric list, homodimeric list and heterodimeric list of protein chains. The hydrophobic fraction is expected to have higher value towards the core of the protein (as indicated by the notion of Hydrophobic Collapse[51]) than towards the protein surface, where its value should be minimum. As

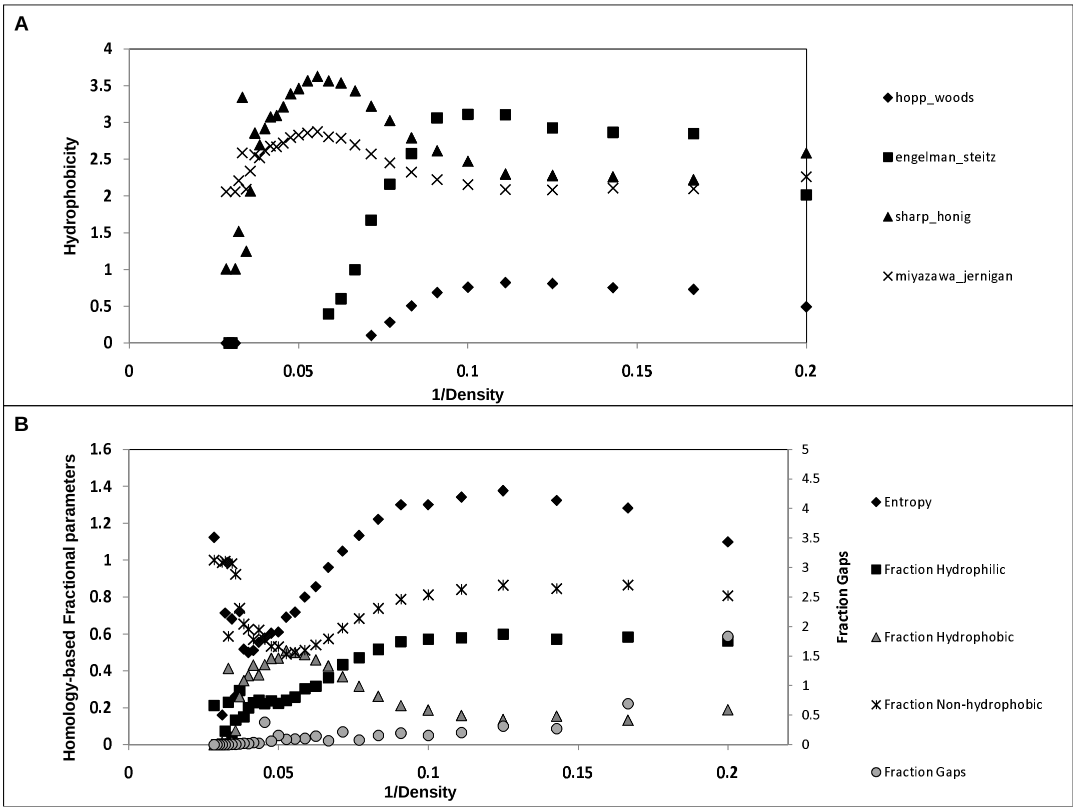


Figure 3.5: Choice of hydrophobicity scale and parameters. Aggregate correlation plots of homology-based parameters averaged over 75891 query residues belonging to 268 learning set list of proteins (ordinate) has been plotted against inverse packing density (abscissa). A total of 235138 alignments were generated for these calculations. A. An overlay plot of aggregate hydrophobicity values obtained from different hydrophobicity scales (ordinate) like, hopp_woods, engelman_steitz, sharp_honig and miyazawa_jernigan, has been plotted against inverse packing density (abscissa). B. An overlay plot of aggregate entropy (ordinate, primary)and other homology based parameters (ordinate, secondary) like, fraction hydrophobic, fraction hydrophilic, fraction non-hydrophobic and fraction gaps versus inverse packing density (abscissa).

expected, the aggregate trend of fraction hydrophobic attains a maxima at lower inverse packing density (higher packing density) and decreases as we move towards higher inverse packing density value (lower packing density) from that maxima. Hence, from the study of the two homology-based parameters, fraction hydrophilic and fraction hydrophobic, it was observed that although a coarse grained approach, the packing density calculation method used in this work, provides a rough representation of surface accessibility. Hence, higher packing density refers to the core of the protein and lower packing density refers to the protein surface.

Again, for the confirmation of this notion, a previously recognized surface accessibility determination method from all atom $3D$ coordinates, NACCESS,[31] was used to compare the packing density values with the relative surface accessibility values. Figure 3.7, represents individual plots of relative surface accessibility as a function of packing density for the four protein lists.

The aggregate plots of relative surface accessibility (RSA) versus packing density shows two major regions, one in which the RSA value decreases linearly with increase in packing density and the other in which the RSA value stays almost constant with increase in packing density. For all the four protein lists, the relative surface accessibility values (RSA) was found to be high at lower packing density values and almost zero at high packing density positions. This suggests that the residues associated with lower packing density have higher relative surface accessibility and hence are expected to lie on the surface of the protein. And, as expected, the residues that are associated with high packing density have almost zero surface accessibility, which means that they form the core of the protein. Hence, it means that the coarse grain approach of calculating packing density used in this work coincides with the NACCESS surface accessibility predictions. Although this is the general trend for most of the protein residues, a few residues associated with lower packing density positions have RSA value equal to zero. Deviations from the aggregate trend shown by

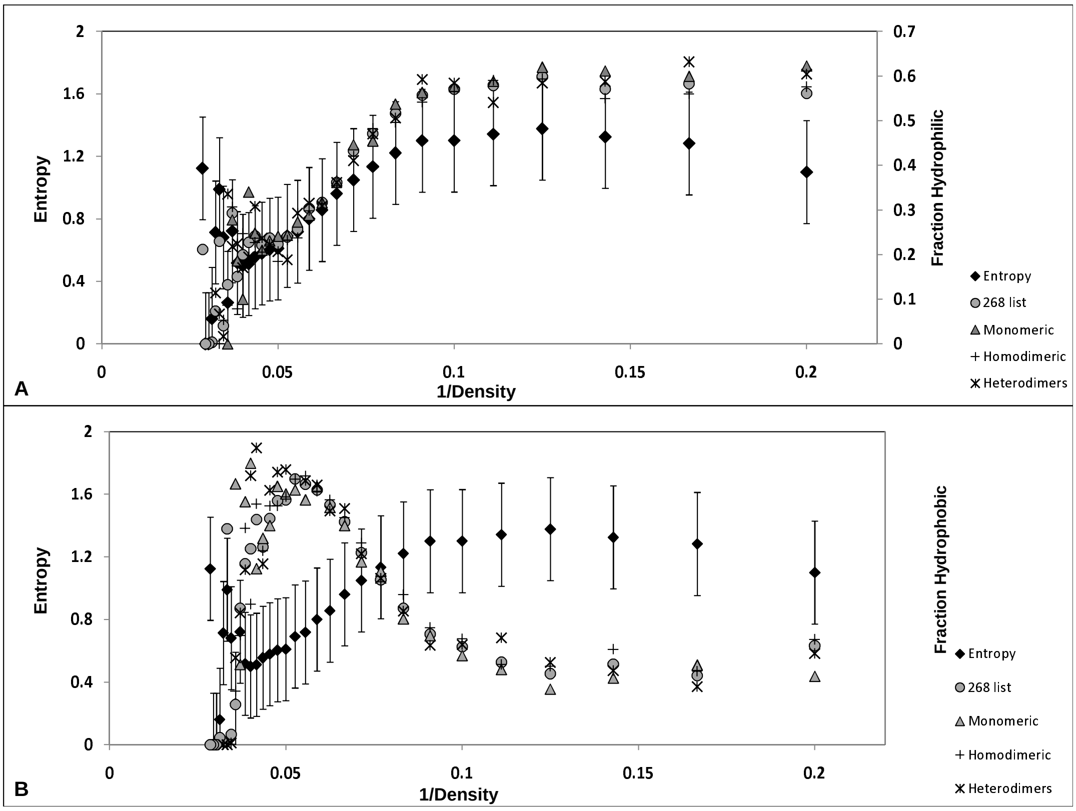


Figure 3.6: Hydrophobic hydrophilic fraction comparison. An aggregate correlation plot of homology based parameters (ordinate) for 268 protein chains from the learning set list, 75 monomeric proteins, 106 homodimeric proteins and 50 heterodimeric protein chains versus inverse packing density (abscissa). A. An aggregate overlay plot of 40% entropy (ordinate, left) of the learning set list and fraction hydrophilic (ordinate, right) for all the four protein lists versus inverse packing density (abscissa). B. An aggregate overlay plot of 40% entropy (ordinate, left) of the learning set list proteins and fraction hydrophobic (ordinate, secondary) for all the four protein lists versus inverse packing density (abscissa).

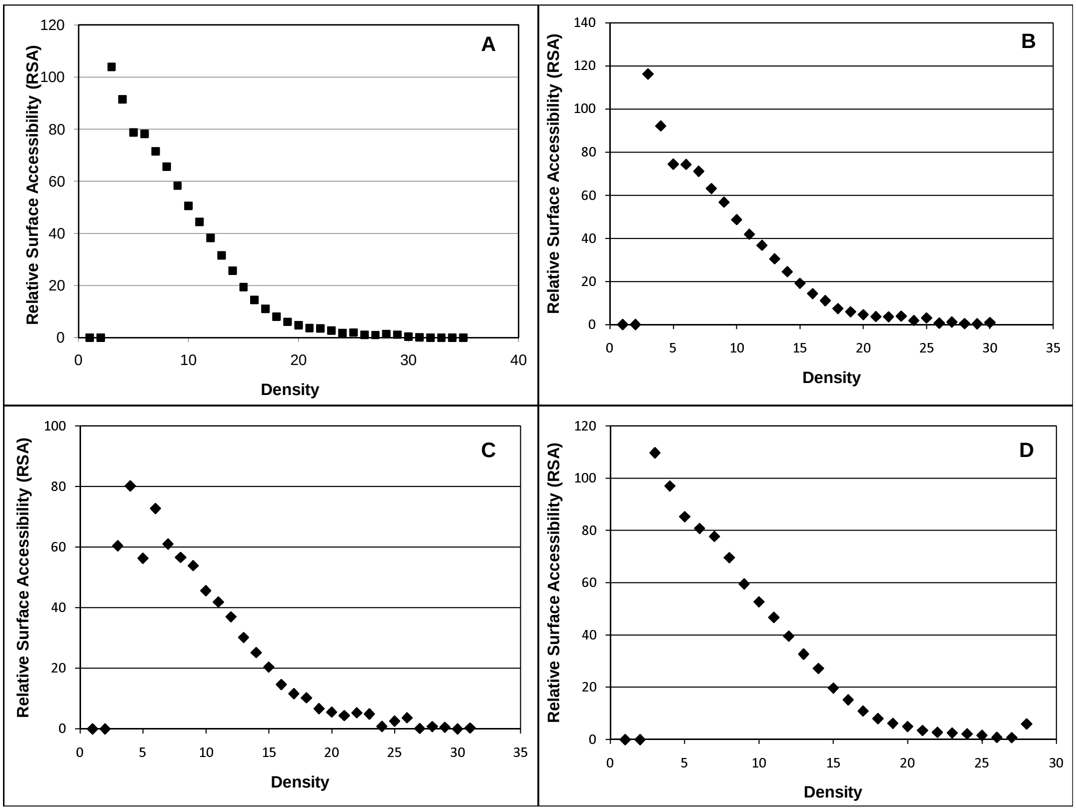


Figure 3.7: Density-relative surface accessibility comparison. Aggregate correlation plot of relative surface accessibility (ordinate) (■, obtained from NACCESS) and packing density (abscissa) for A. 75 specially characterized monomeric proteins. Here the aggregate RSA values were obtained by double averaging a total of 20076 query residues at each packing density position. B. 106 homodimeric proteins. Here, the RSA values of a total of 28734 query residues were double averaged to obtain the aggregate RSA values. C. 50 heterodimeric list. Here, the aggregate RSA values were obtained by averaging RSA values from a total of 9665 query residues and D. 268 learning set list of proteins. The RSA values were obtained by double averaging relative surface accessibility values for a total of 75891 residues at each packing density position.

the learning set list of proteins (Figure 3.7D) is more pronounced for the heterodimeric aggregate list (Figure 3.7C). While the monomeric list shows minimal deviations from the two region trend, aggregate analysis with the homodimeric list (Figure 3.7B) shows an intermediate behavior.

3.2.2 Validation of Entropy Calculations

Various entropy values calculated from different methods suggested in the literature have been evaluated here. The entropy values were calculated by taking into account only the residues but not the gaps. Three types of entropy values, 40% entropy, 6-point entropy and HSSP derived entropy values (as discussed in Chapter 2) were calculated and compared against each other. An overlay plot of the three entropy values for the 268 protein list is shown in Figure 3.8.

From the entropy- inverse density plot shown in Figure 3.8, the nature of the curve is found to be similar in shape for all the three entropy values. This proves that the method of entropy calculation used in this study gives comparable outcomes as any other prevalent entropy calculation methods. From this aggregate plot, two regions have been specified, Region I that increases linearly with inverse density and Region II that stays more or less the same with increase in inverse density.

3.2.3 Entropy Inverse-Density Correlation Plots

The aggregate trend of the entropy-inverse density plot for the learning set list, monomeric list, homodimeric list and heterodimeric list is shown in Figure 3.9. The aggregate entropy-inverse density correlation plots of the four lists show similar trends. Two major regions were consistently noted in all these graphs. Major Region I, associated with packing density $11-25$, corresponds to the portion of the curve where average sequence entropy increases linearly with increase in packing density. Major Region II, associated with the packing density $4-10$, corresponds to the portion where

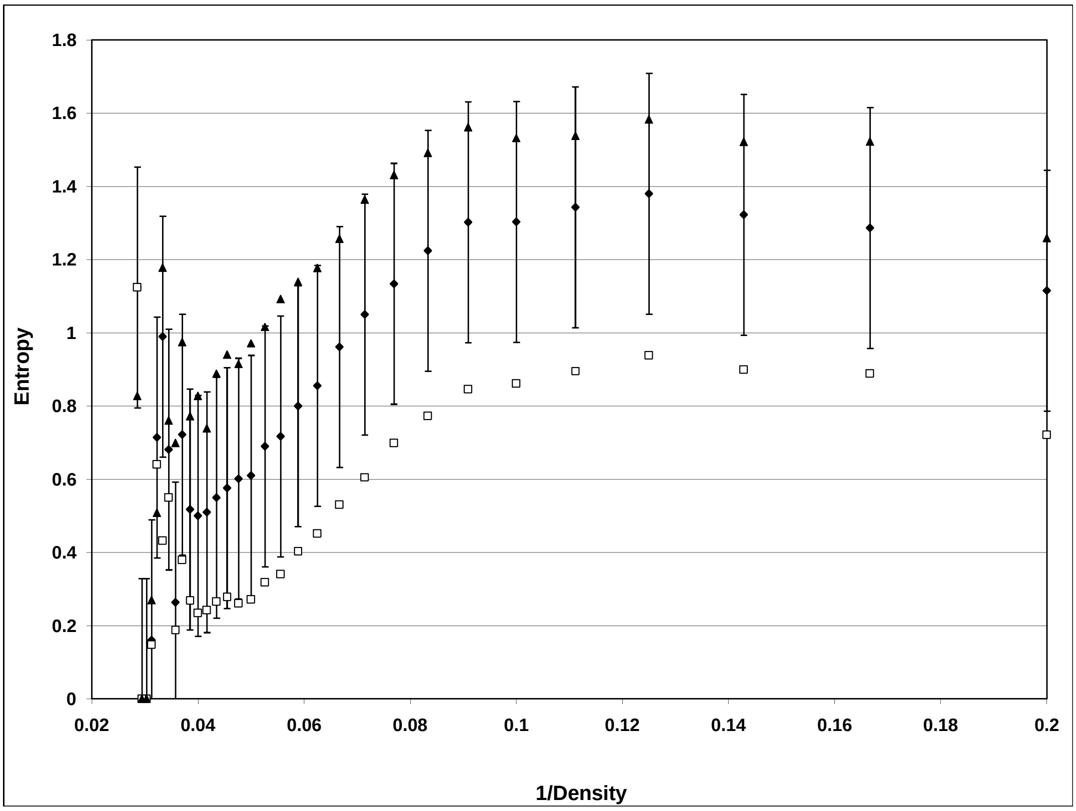


Figure 3.8: Various entropy comparison. A comparison plot of average sequence entropies for the learning set list of 268 proteins: Gaps-excluded standard entropy (♦), 6− term gaps-excluded entropy (□), HSSP-derived entropy (▲). The various average sequence entropy values (ordinate-left) corresponding to 75891 query residues of the learning set list of 268 proteins, are calculated by double averaging the entropy values at each inverse packing density position (abscissa).

sequence entropy remains almost the same with increase in packing density. From Z-score analysis, (see Appendix Figure B.5 and Table C.4) the residues with packing density less than 4 and packing density greater than 25 has been characterized as falling in the anomalous region (consisting of approximately 0.31% of the total residues) of the entropy-inverse density plot. The corresponding population for sequence entropy involves a small fraction of some 0.25% of query residues at the corresponding packing densities greater than 25. However, the fraction of query residues with C_α packing region greater than twenty was found to be 6.64%. The fraction of query residues with packing densities less than four is 0.05%.

All the three plots obtained from the monomeric, homodimeric and heterodimeric lists were compared to the aggregate learning set list plot as this plot is a mixture of all the aforementioned other three list types (see Appendix Table C.3, C.1 and C.2). The average ratio of the average entropy values of the learning set list with the average entropy values for the monomeric, homodimeric and heterodimeric list at each packing density position was found to be 0.996, 1.05 and 1.21. The deviation of these ratios from ratio equal to 1 (measure of absolute similarity), for the monomeric, homodimeric and heterodimeric list was found to be 0.004, 0.054 and 0.211 respectively. Hence, it can be concluded that the trends obtained from the monomeric list and the homodimeric list were found to be closer to the aggregate correlation plot than that obtained from the heterodimeric list.

The correlation of fraction strongly hydrophobic (ordinate, right) with inverse packing density has been depicted in Figure 3.10 in conjunction with aggregate sequence entropy trend (ordinate, left). The curves obtained from the learning set list (Figure 3.10A), monomeric list (Figure 3.10B), homodimeric list (Figure 3.10C) and heterodimeric (Figure 3.10D) list are similar in character. As we move from left to right, the fraction of strongly hydrophobic residues first increases and then attains a maxima at inverse packing density 0.05 after which it starts to decrease with increase

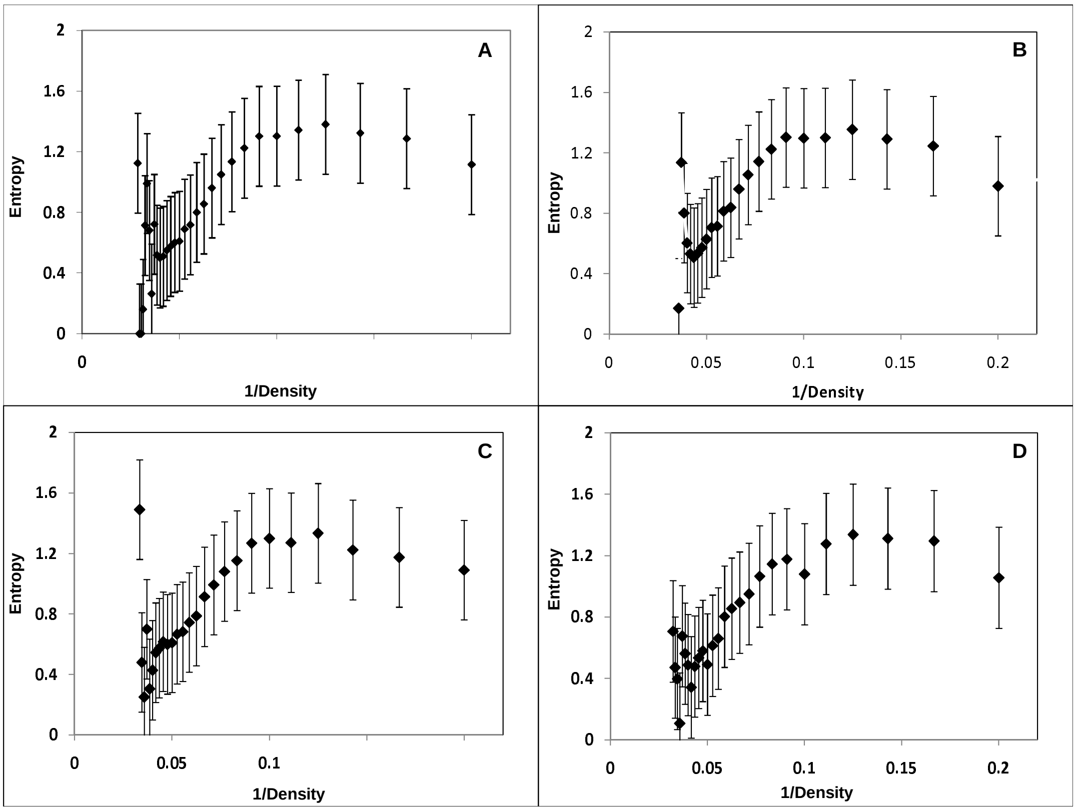


Figure 3.9: Aggregate entropy-inverse density trend comparison for the four protein lists. Combined aggregate correlation plots of 40% sequence entropy (ordinate) calculated with gaps excluded versus inverse packing density (abscissa) for A. learning set list proteins. Average standard sequence entropy (♦, left ordinate) has been calculated by averaging the entropy values of 75891 query residues aligned to a total of 235138 subject sequences, at each inverse packing density value (abscissa). B. Monomeric list proteins. Average standard sequence entropy (♦, left ordinate) has been calculated by averaging the entropy values of 20076 query residues aligned to a total of 59751 subject sequences, at each inverse packing density value (abscissa). C. Homodimeric list proteins. Average standard sequence entropy (♦, left ordinate) has been calculated by averaging the entropy values of 28734 query residues aligned to a total of 83048 subject sequences, at each inverse packing density value (abscissa). and D. Heterodimeric list. Average standard sequence entropy (♦, left ordinate) has been calculated by averaging the entropy values of 9665 query residues aligned to a total of 55683 subject sequences, at each inverse packing density value (abscissa).

in inverse packing density.

Two types of residues are present in major Region I, the first type of residues have exclusively strongly hydrophobic residues in their alignments and the second type of residues have the presence of some other type of residues, like small residues or non-strongly hydrophobic residues, in addition to the strongly hydrophobic residues in their aligned residues. The aggregate plot of fraction strongly hydrophobic versus inverse density, for both type of residues show similar trends (Appendix Figure B.6), a noticeable peak between inverse packing density $0.04 - 0.06$.

The combined aggregate correlation plots of the fraction of residues that are strongly hydrophobic (Figure 3.10 at each aligned residue position of a protein), show a sharp peak at packing density 18 (Full Width at Half Maximum is 0.228). It means that the highest value of fraction strongly hydrophobic $(FSHP = 0.45)$ is at inverse packing density 0.056. Hence, at least 45% of the proteins (a sum of both both query (268 protein chains) and subject proteins (235138 proteins)) have strongly hydrophobic residues at that inverse packing density position. Since this peak is very sharp, most of the strongly hydrophobic residues, belonging to a total of 268 protein chains aligned to 235138 subject proteins with a total of $7.12E7$ residues, are expected to have their densities centered around packing density 18. Hence, it can be inferred that the formation of a strongly hydrophobic core[51] might occur at or around this packing density. Hence, substitution with strongly hydrophobic residues should be clustered around this packing density point.

Now, approximately, 24.70% of the aligned residues that are strongly hydrophobic are associated with query sequence positions at which an exclusive presence of strongly hydrophobic residues (FSHP $= 1$) has been noticed. From this percentage, 68.20% of the aligned residues (with FSHP $= 1$) are present between packing densities $14 - 22$ (shown in Appendix Figure B.6). The mean of the residues with $FSHP = 1$, lies between packing density $14 - 19$. Hence, it provides additional support to the conjecture that the strongly hydrophobic core forms at a critical distance from the densest portion of the protein (somewhere between packing densities $14-22$), and is centered around packing density 18. From all this, the presence of a critical percentage (around 70%) of aligned strongly hydrophobic residues, involved in the hydrophobic collapse is indicated. From the aggregate plots and the frequency distributions for FSHP at various density positions, a total absence of strongly hydrophobic residues beyond the volume of strongly hydrophobic core is not observed.

The aggregate learning set list plot of fraction strongly hydrophobic as a function of inverse packing density, was compared with the trends of fraction strongly hydrophobic versus packing density for the monomeric (Figure 3.10B), homodimeric (Figure 3.10C) and heterodimeric (Figure 3.10D) lists. The average ratio of fraction strongly hydrophobic (FSHP), from the individual ratios of FSHP from learning set list and each of the monomeric list, homodimeric list and heterodimeric list, at each density position was found to be 1.065, 0.963 and 1.109, respectively. The deviation of these ratio from 1 (measure of absolute similarity) was found to be 0.065, 0.037 and 0.109 for the monomeric, homodimeric and heterodimeric lists, respectively. Hence, it is observed that the aggregate fraction strongly hydrophobic curve obtained from the heterodimers have the largest deviations from the aggregate learning set list plots. Although, the deviations of monomeric and homodimeric protein lists are of the same order of magnitude, homodimers have lesser deviations from the learning set list than the monomers.

Figure 3.11 represents the overlay plot of entropy density (ordinate, left) and fraction small residues (ordinate, right) as a function of inverse packing density for the learning set list proteins (Figure 3.11A), the monomeric protein list (Figure 3.11B), the homodimeric protein list (Figure 3.11C) and the heterodimeric protein list (Figure 3.11D). As we move from left to right, the fraction of small residues decreases with increase in inverse packing density and attains a minima at inverse packing density

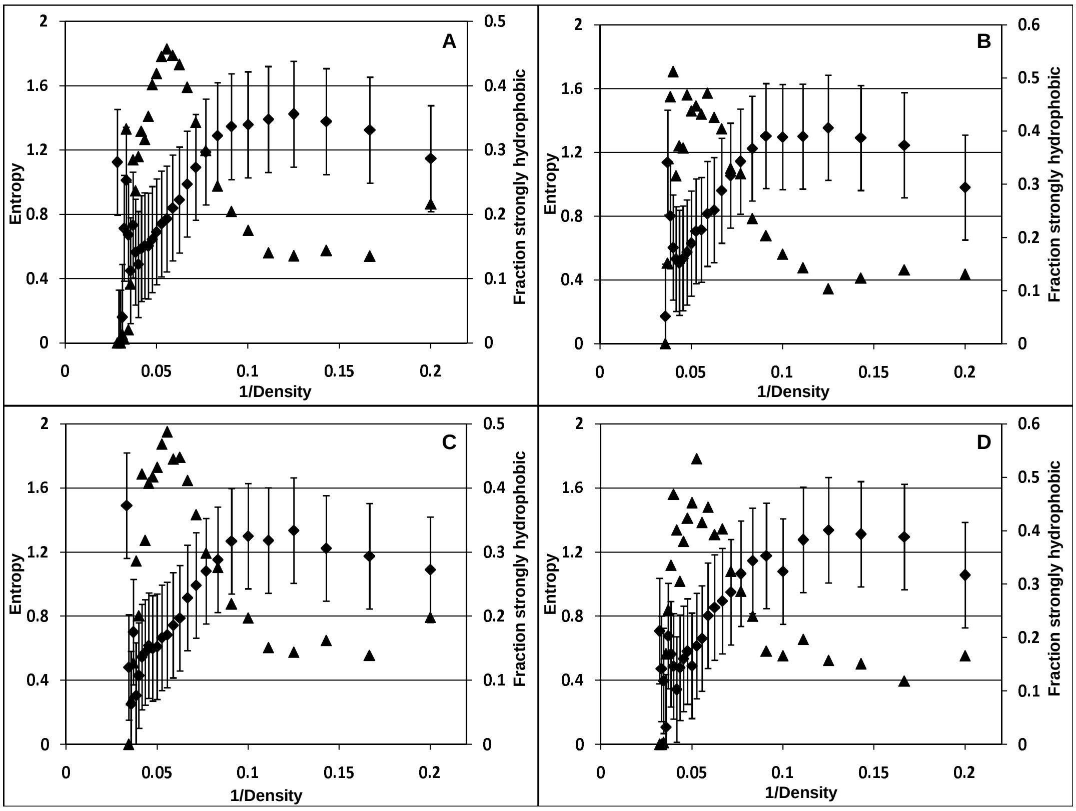


Figure 3.10: Aggregate fraction strongly hydrophobic-inverse density trend comparison for the four protein lists. Aggregate overlay correlation plots of entropy (♦, ordinate left) and fraction strongly hydrophobic (▲, ordinate right) as a function of inverse packing density (abscissa) for A. 268 learning set list of proteins, B. 75 monomeric proteins, C. 106 homodimeric proteins and D. 50 heterodimeric list. Average standard sequence entropy and average fraction strongly hydrophobic have been calculated by double averaging their values from 75891 query residues aligned to 235138 subject sequences for the learning set list proteins, 20076 query residues aligned to 59751 subject sequences for monomeric proteins, 28734 query residues aligned to 83048 subject sequences for homodimeric proteins and 9665 query residues aligned to 55683 subject sequences for the heterodimeric list at each inverse packing density value (abscissa).

0.06 after which the fraction of small residues increases slightly and stays more or less the same with increase in inverse packing density.

The aggregate plot of fraction small residues versus inverse density, for the learning set list, shown in Figure 3.11, exhibits a transient inflexion between packing density $14 - 22$. Approximately 40.32% of aligned small residues are present between this packing density. At packing density 14, 12.8% of aligned small residues are present. As we move left from this point this percentage of aligned small residues increases first gradually and then sharply with decrease in inverse packing density. And, from this point the percentage of aligned small residues first increases and then stays more or less constant with increase in inverse packing density. All this suggests that the region of strongly hydrophobic core is marked with a minimum presence of small residues. Since, all these packing densities discussed above lie in the major Region I, the residues falling in this region should be buried and the residues falling in the major Region II should be marked as surface accessible.

The plot of fraction of small residues (FSR) versus inverse packing density for the learning set list of proteins has been compared to that of the monomeric, homodimeric and heterodimeric lists. The average value of the ratios of aggregate fraction small residue versus inverse packing density plot of learning set list to monomeric list, homodimeric list and heterodimeric list, calculated at each inverse packing density position was found to be 1.053, 0.988 and 1.463, respectively. This suggests that the deviations of the FSR-inverse density plot obtained from monomeric list, homodimeric list and heterodimeric list are 0.053, 0.012 and 0.463, respectively. As is evident from these values, the plots obtained from the heterodimeric protein list show the largest deviations from the mixed aggregate plot of learning set list of proteins.

The overlay of entropy and fraction of non-strongly hydrophobic (FNSHP) residue trends as a function of inverse packing density has been plotted in Figure 3.12 for the learning set list (Figure 3.12A), monomeric list (Figure 3.12B), homodimeric list

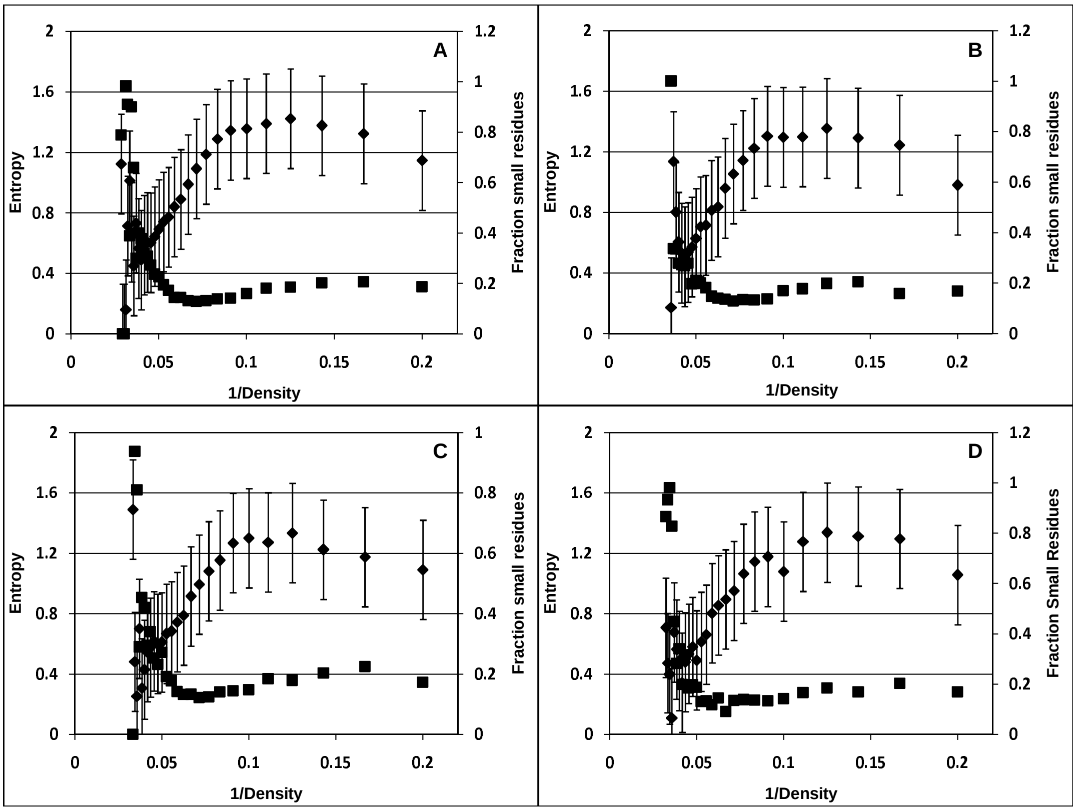


Figure 3.11: Aggregate fraction small residues-inverse density trend comparison for the four protein lists. Aggregate overlay correlation plots of entropy (♦, ordinate left) and fraction small residues (■, ordinate right) as a function of inverse packing density for the list of A. 268 learning set list of proteins, B. 75 monomeric proteins, C. 106 homodimeric proteins and D. 50 heterodimeric protein list. Average standard sequence entropy and average fraction small residues have been calculated by double averaging their values from 75891 query residues for the learning set list proteins, 20076 query residues for the monomeric proteins, 28734 query residues for the homodimeric proteins and 9665 query residues for the heterodimeric list at each inverse packing density value (abscissa).

(Figure 3.12C) and the heterodimeric list (Figure 3.12D) of proteins. The aggregate trend of FNSHP suggests a decrease in the value of FNSHP with increase in inverse packing density until the point of minima (0.05 inverse packing density) that is observed in major Region I. After this minima, FNSHP increases with increase in inverse packing density (till packing density 0.10) forming a parabola type of shape, after which the value of FNSHP remains more or less constant with further increase in packing density.

The aggregate plot of fraction non-strongly hydrophobic versus inverse packing density, for the learning set list, shown in Figure 3.12, attains a minima between packing density $14 - 22$ that is centered at packing density 18. The minima suggests the fact that at least 54.28% of the non-strongly hydrophobic residues are present at packing density 18. Approximately 42.95% non-strongly hydrophobic residues are present between this packing density. This means that compared to any other density positions, around packing density 18 (that is between densities $14 - 22$), the presence of FNSHP residues is the least. Hence, the density range $14 - 22$, is dominated by the presence of strongly hydrophobic residues whose relative composition with fraction non-strongly hydrophobic changes as we move away from the center of the strongly hydrophobic core.

The FNSHP plots from learning set list was compared to those obtained from monomeric, homodimeric and heterodimeric lists. The average value of the ratios obtained from the aggregate plot of FNSHP versus inverse packing density of learning set list and monomeric, homodimeric and heterodimeric lists were found to be 1.008, 0.999, and 1.024, respectively. The deviation of these values from ratio equal to 1.0 (measure of absolute similarity) was found to be 0.008, 0.001 and 0.024, respectively. The deviation of FNSHP versus inverse packing density curve for heterodimers was found to have maximum deviation and hence maximum dissimilarity from the mixed aggregate plot of learning set list for the same parameter than the other two lists.

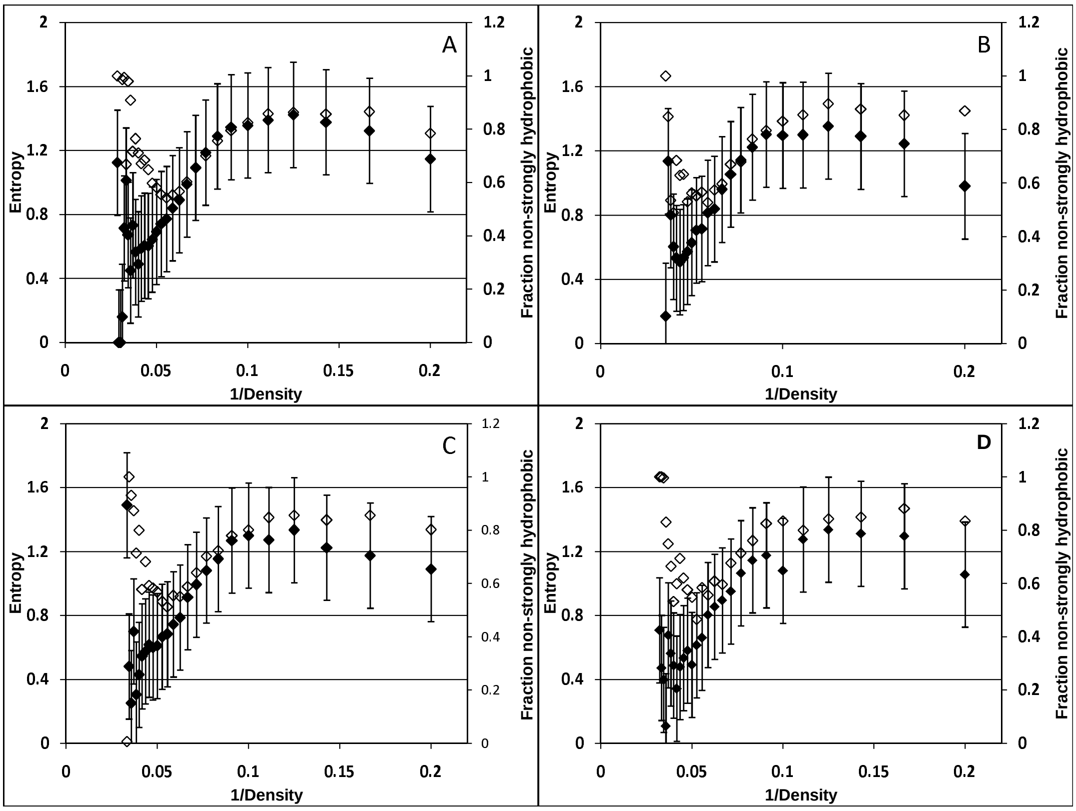


Figure 3.12: Aggregate fraction non-strongly hydrophobic vs inverse packing density trend comparison for the four protein lists. Aggregate overlay correlation plots of entropy (♦, ordinate left) and fraction non-strongly hydrophobic (◊, secondary axis) as a function of inverse packing density (abscissa) for A. 268 Learning set list of proteins, B. 75 monomeric proteins, C. 106 homodimeric proteins and D. 50 heterodimeric protein list. Average standard sequence entropy and average fraction non-strongly hydrophobic residues have been calculated by double averaging their values from 75891 query residues for the learning set list proteins, 20076 query residues for the monomeric proteins, 28734 query residues for the homodimeric proteins and 9665 query residues for the heterodimeric list at each inverse packing density value (abscissa).

The aggregate overlay plot of entropy (ordinate, left) and fraction of gaps (ordinate, right) versus inverse packing density has been plotted in Figure 3.13, for the learning set list (Figure 3.13A), the monomeric list (Figure 3.13B), the homodimeric list (Figure 3.13C) and the heterodimeric list (Figure 3.13D). The average value of the fraction of gaps has been found to increase steadily with increase in packing density. A total of 43.16% residues lying between packing density $14 - 22$ have gaps in their alignments. A comparison of this aggregate curve has been carried out for all the four lists. The average value of the ratios obtained by dividing fractional values of learning set list from the fractional values of the monomeric, homodimeric and heterodimeric lists at each packing density position were found to be 1.03, 1.91 and 4.18 respectively. The deviations of these average ratios from 1.00 (measure of absolute similarity), were found to be 0.03, 0.91 and 3.18 for the monomeric, homodimeric and heterodimeric lists respectively. Here also, the aggregate plot obtained from the heterodimeric protein list was found to be the most dissimilar to the mixed aggregate plot obtained from the learning set list.

Figure 3.14 represents the aggregate overlay plot of the various homology-based parameters for the four protein lists, namely, learning set list (Figure 3.14A), monomeric list (Figure 3.14B), homodimeric list (Figure 3.14C) and heterodimeric list (Figure 3.14D) from the entropy-inverse density plot, two major regions, namely Region I and Region II, have been indicated. The average value of fraction strongly hydrophobic attains maxima (corresponding to inverse packing density $0.05 - 0.06$) in major Region I and has an overall shape of a parabola. The aggregate curve of fraction non-strongly hydrophobic attains a minima (corresponding to inverse packing density $0.05 - 0.06$) in the major Region I and looks like mirror image of the curve obtained from fraction strongly hydrophobic versus inverse packing density. The inflexion point of the aggregate average plot of fraction of small residues as a function of inverse packing density was found to be in major Region I (associated with an inverse packing density value of 0.06). The aggregate trend of fraction gaps increases uniformly from lower inverse packing density value region to higher inverse packing density value regions. The aggregate trend of small residues (AG) resembles the non-strongly hydrophobic trend more than the strongly hydrophobic trend. Also, despite being non-polar in character, small residues could not be classified as strongly hydrophobic

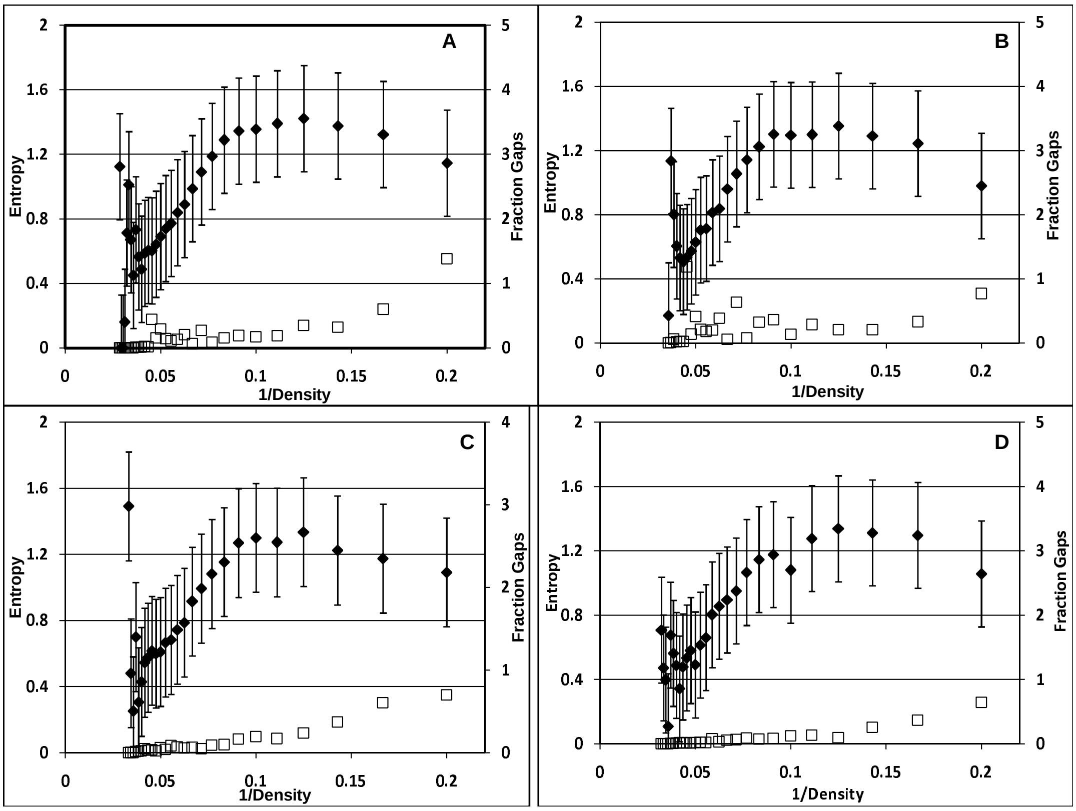


Figure 3.13: Aggregate fraction gaps-inverse density trend comparison for the four protein lists. Aggregate overlay correlation plots of entropy (♦, ordinate left) and fraction gaps (fraction gaps (□, ordinate right) as a function of inverse density (abscissa) for A. 268 Learning set list of proteins, B. 75 monomeric proteins, C. 106 homodimeric proteins and D. 50 heterodimeric protein list. Average standard sequence entropy and average fraction gaps have been calculated by double averaging their values from 75891 query residues for the learning set list proteins, 20076 query residues for monomeric proteins,28734 query residues for homodimeric proteins and 9665 query residues for the heterodimeric list at each inverse packing density value (abscissa).

residues.[32,52]

From a subjective comparison of the various homology-based parameter plots for the monomeric, homodimeric and heterodimeric lists, the plots obtained for the heterodimeric list proteins appear to have the most deviations from the aggregate trends of the learning set.

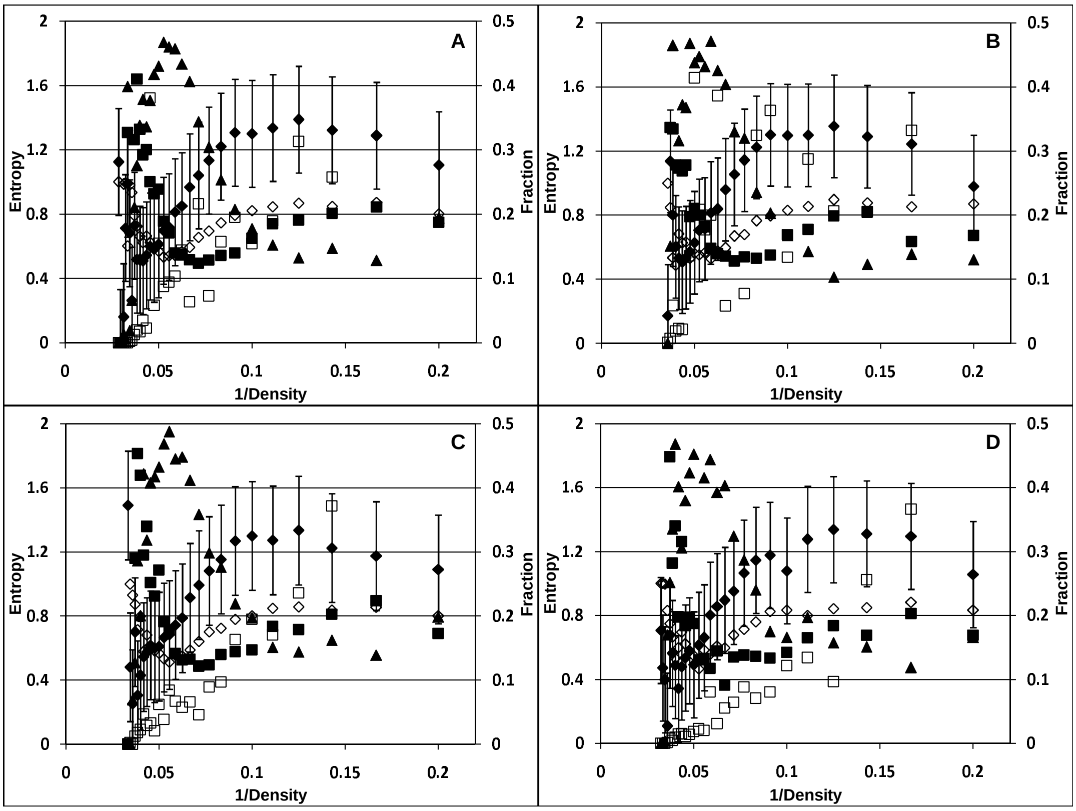


Figure 3.14: Aggregate correlation plots of various homology-based parameters and inverse packing density. An aggregate overlay plot of entropy(♦, ordinate left), Fraction non-strongly hydrophobic (◊, ordinate left) Fraction gaps (□, ordinate right), Fraction small residues (■, ordinate right), and Fraction strongly hydrophobic (▲, ordinate right) as a function of inverse packing density (abscissa) for A. Learning set proteins, B. monomeric proteins, C. homodimeric proteins and D. heterodimeric protein list. The aggregate values of all the homology-based parameters have been calculated by double averaging a total of 75891 query residues for the learning set proteins, 20076 query residues for the monomeric protein list, 28734 query residues for the homodimeric protein list and 9665 query residues for the heterodimeric protein list.

3.3 Frequency Distributions

In this section, the frequency of residues associated with density and various homology-based parameters have been evaluated. A total of 75891 aligned residues belonging to 268 protein chains of the learning set list, have been processed for these plots. Due to the non-availability of either the coordinate information from the mmCIF files, or the BLASTP alignment results for any particular residue position, 2121 aligned residues were discarded. Also, a total of 43 residue positions, flagged due to the presence of only gaps, were also eliminated leaving a total of 73727 residues to be used for frequency distribution plots. Two type of distributions, first based on the density of the homology-based parameters and second on the basis of the value of the homology-based parameters, have been analyzed in this section.

3.3.1 Density Distribution

In Figure 3.15, the frequency of strongly hydrophobic residues (Figure 3.15A), frequency of small residues (Figure 3.15B), frequency of non-strongly hydrophobic residues (Figure 3.15C) and frequency of fraction gaps (Figure 3.15D), for the learning set list, has been distributed over appropriate packing density ranges. The distribution of residues that are strongly hydrophobic over each density bin looks like a normal distribution that is slightly left skewed at the top. The distribution attains a maxima at a packing density of $15 - 16$ (that is $0.067 - 0.063$ inverse packing density). Approximately 4.01% of strongly hydrophobic residues are present in the packing density range at which the aggregate curve of fraction strongly hydrophobic vs. inverse packing density attains a maxima at 0.45. The small residue distribution over the density bins represent left skewed normal distribution. The frequency of small residues is maximum between density positions 11 and 12 where approximately 18.36% residues are concentrated. Only 3.81% small residues are present at the packing density range where the inflexion point of the aggregate plot of fraction of small residues plotted as a function of inverse packing density, is observed. The density distribution of the non-strongly hydrophobic residues represents a distribution that is slightly flattened at the top. A nominal right skewedness of the graph is also observed. Approximately, 43.45% of the non-strongly hydrophobic residues are present at the top of the distribution. The inflexion point observed in the aggregate plot of the fraction non-strongly hydrophobic plotted as a function of inverse packing density corresponds to packing density $19 - 20$ at which position, approximately 20.46% non-strongly hydrophobic residues are present. The distribution obtained for fraction of gaps looks like a normal distribution with a little dip at the top of the curve between $12 - 13$ packing density. The maximum value attained by the histogram corresponds to packing density 15 (inverse packing density 0.06) and 11 (inverse packing density 0.09).

An overlay plot of packing density (Figure 3.16A) and inverse packing density (Figure 3.16B) distributions for the homology-based parameters of the learning set list of proteins, has been shown in Figure 3.16. Although the total number of strongly hydrophobic residues is more than small residues, the nature of the histogram for small residues and strongly hydrophobic residues appear to be complementary in nature. The number of residues that are non-strongly hydrophobic is the largest followed by the strongly hydrophobic residues. The total number of residues that are small have the smallest occupancies when compared to the other distributions.

3.3.2 Bin Distribution

Here, the frequency of residues, with respect to the values of the homology-based parameters has been shown for the learning set list (Figure 3.17). The ratio of total number of residues present in major Region I to major Region II is approximately 4.58, which means major Region I is highly populated. The distribution of residues with respect to entropy values (Figure 3.17A), shows a maxima, for both major Region

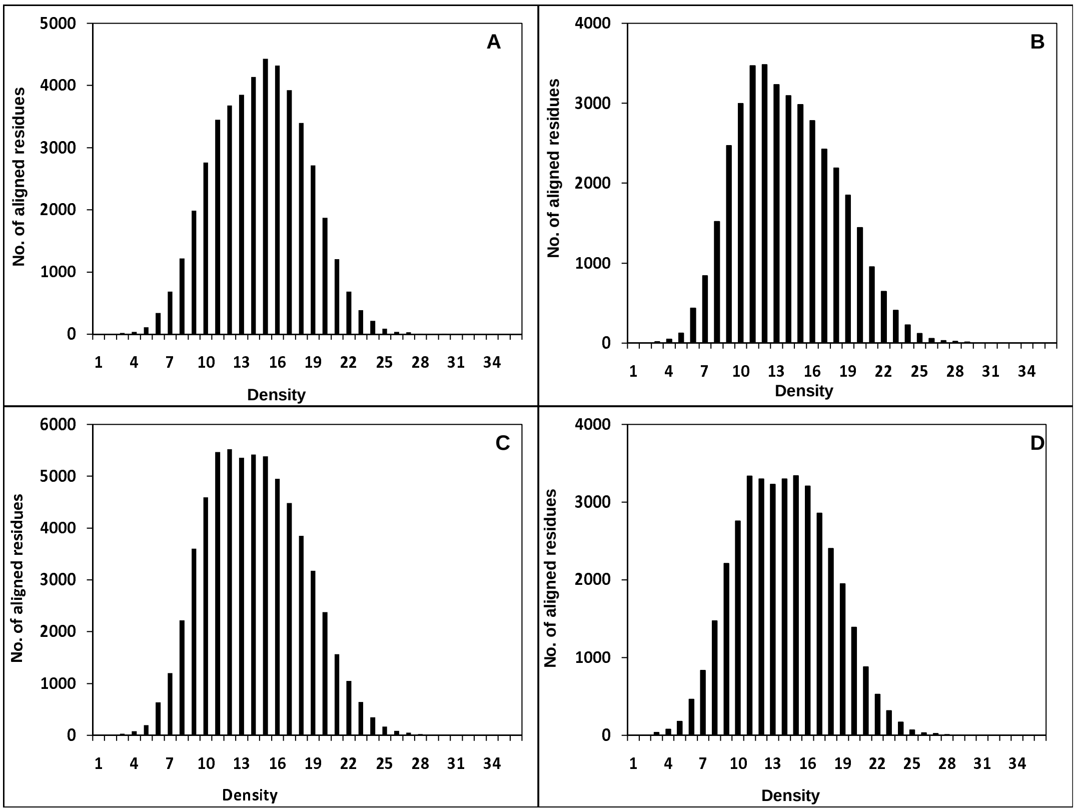


Figure 3.15: Density distribution plots for the fractional parameters. Frequency distribution plots for A. Fraction strongly hydrophobic, B. Fraction small residues, C. Fraction non-strongly hydrophobic and D. Fraction Gaps. Here, frequency of 73727 query residues (ordinate) belonging to a total of 268 protein chains have been calculated at each density position (abscissa).

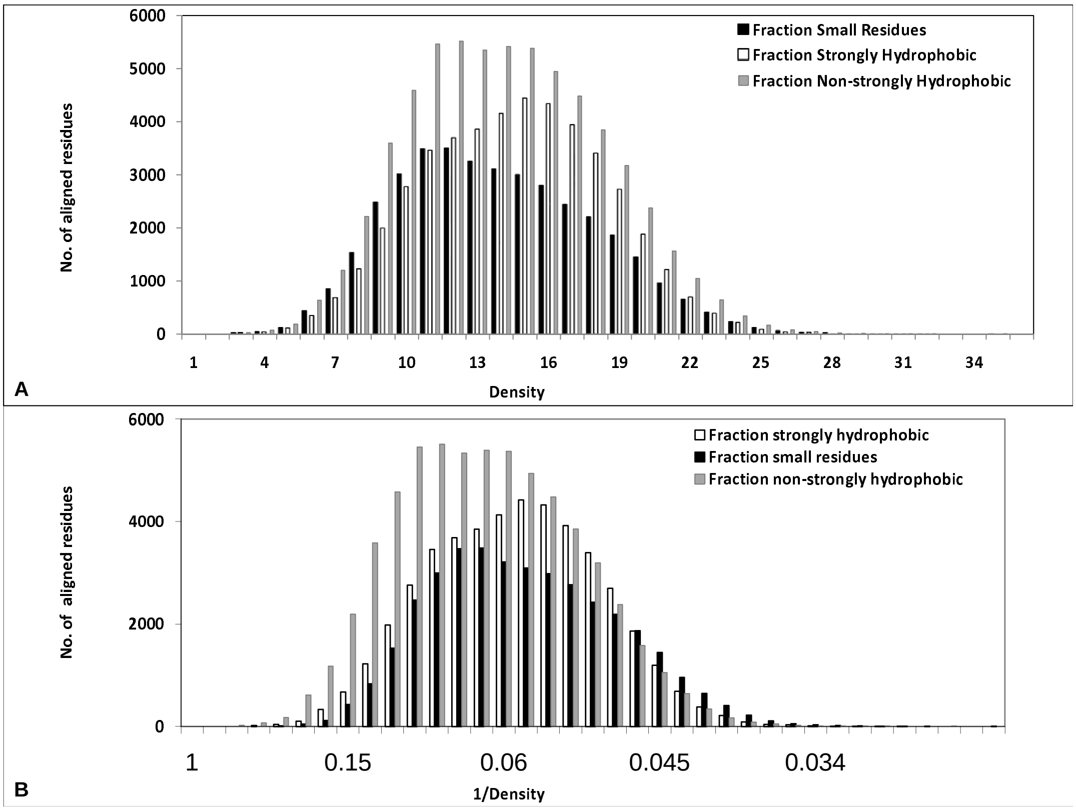


Figure 3.16: Overlay plot of density distributions for various fractional parameters. Frequency distributions (ordinate) are calculated for a total of 73727 query residues at each density position (abscissa). A. An overlay frequency distribution plot of 268 protein chains for a set of homology based parameters (ordinate) at each density bin (abscissa). B. An overlay frequency distribution plot of homology based parameters (ordinate) for a set of 268 protein chains plotted against inverse packing density bin (abscissa).

I and major Region II at entropy value 0. Hence, 19.79% of residues belonging to major Region I and 13.01% of residues belonging to major Region II are highly conserved. For major Region I, the frequency of residues first decreases, then stays almost flat for a while, then again attains a maxima and then decreases linearly as density increases. An increase in entropy means an increase in partitioning between the amino acids. The minima in the curve is observed at entropy bin $0.6 - 0.8$. This corresponds approximately to packing density $16 - 21$. And this is the density position at which the fraction strongly hydrophobic attains a maxima in the inverse density plot. This indicates that there is a certain fraction of residues (7.61%) and a particular partitioning (entropy $0.6 - 0.8$) associated with the packing density at which the fraction of strongly hydrophobic residues is maximum. Major Region II trends remains almost flat after the first entropy equal to 0 peak. Hence, the effect of partitioning is meaningless in this case. For both major Region I and major Region II, equi-partitioning of residues at any sequence position is a highly unlikely event because at higher entropy values relatively fewer residues are observed. The ratio of occupancy of the two regions (plotted in the insert), exhibit a linear decrease with increase in the entropy value. Hence, as we move from left to right in Figure 3.17, this linear decrease in ratio (R= 0.953, slope= -1.497), suggests an associated increase in the predominance of surface accessible Region II. Since the strongly hydrophobic core is expected to occur in major Region I (buried region of the protein), the inflexion point of homology-based parameters can serve as filtering thresholds for surface accessibility prediction.

The frequency distribution of fraction strongly hydrophobic (FSHP), for the learning set list of proteins, plotted in Figure 3.17B, shows a bimodal character for the two regions. Residues having 0 fractional values indicate an absence of strongly hydrophobic residues which is seen in 22100 aligned residues for major Region I and 6072 aligned residues for major Region II. After eliminating the data points associated with a fractional value of 0, the ratio of major Region I and major Region II residue frequency was found to be 5.37. The filter threshold for fraction strongly hydrophobic lies at the inflexion point of the aggregate fraction strongly hydrophobic-inverse density plot and also lies at the middle point of the bimodal distribution bifurcating the two distributions. This conservatively chosen filter demarcates the presence of 50336 residues (68.27% of aligned residues) as surface exposed (FSHP < 0.5) and 23391 residues (31.73% of aligned residues) as buried residues. The ratio of the residues present in the two halves, FSHP < 0.5 and FSHP ≥ 0.5, was found to be 2.15. In Figure 3.17B, it was observed that the ratio of major Region I to major Region II increases linearly (R= 0.849, slope= 7.902) as a function of FSHP. This further reinforces the notion that residues with FSHP value ≥ 0.5 should be associated with a higher probability of being characterized as buried, since major Region I predominates here at each value higher than 0.5.

The frequency distribution of fraction of small residues, for the learning set list of proteins, has been shown in Figure 3.17C. Fraction of small residues (7.4% of the total processed residues) has considerable number of residues (7.85% in Region I and 5.37% of Region II) with value 1. This means that the proteins have some portions/regions where there is absolute presence of small residues. And this portion/region is present in case of both region I and region II. This means that small residues are found in the core as well as on the surface of the proteins.

In Figure 3.17C, the frequency of residues present at fraction 0 means an absence of small residues in 31099 aligned residues belonging to Region I and 4780 aligned residues of Region II. After eliminating this fraction, a bimodal frequency distribution curve is observed. The ratio of frequency of aligned small residues present in Region I and Region II was found to be 3.48. The filter threshold for fraction small residues lies at the point of inflexion observed in the fractional parameters-inverse packing density curve. This conservatively chosen point of fraction small residues, < 0.15, is associated with significant population on both side of it. A total of $56,538$ residues (76.69% of the total residues) and $17,189$ residues(23.31% of the total residues) are present with FSR value of < 0.15 and ≥ 0.15 respectively. The ratio of residues of these two parts, 3.28, is comparable to the ratio of major Region I to major Region II. Also, in Figure 3.17C, a slight linear increase (R = 0.184, slope = 0.902) in the ratio of Region I to Region II is observed as a function of FSR. This predominant presence of Region I associated with fraction small residue value ≥ 0.15 supports the conviction that probability of residues being buried should be greater for residues falling in this half than in the other half (FSR < 0.15).

The frequency distribution of fraction gaps over packing density, for the learning set list of proteins, has been shown in Figure 3.17D. An absence of absolute mutational regions is seen here. The curve is not bimodal and has only one distribution. The frequency distribution of fraction gaps appears more like a left-truncated normal distribution that has little information associated with it in terms of surface accessibility prediction.

Frequency distribution of query residues as a function of entropy values has been shown in Figure 3.18A. An overlay plot of the residue frequencies for a few homology-based parameters like fraction strongly hydrophobic, fraction of residues that are small, and fraction of gaps as a function of packing density, for the learning set list, has been shown in Figure 3.18B. Figure 3.18A is characterized by two maxima and can be thought as a plot of two coupled distributions. In Figure 3.18B, as we move from left to right, at any fractional value, the number of small residues is more than the number of strongly hydrophobic residues until fractional value 0.50, at which point the frequency of small residues and strongly hydrophobic residues is exactly the same. After this point, the ratio of the number of small residues to strongly hydrophobic residues decreases as we further move from 0.50 to 1.00. Hence from this comparison it is observed that the relative ratio of the small residues and strongly hydrophobic residues

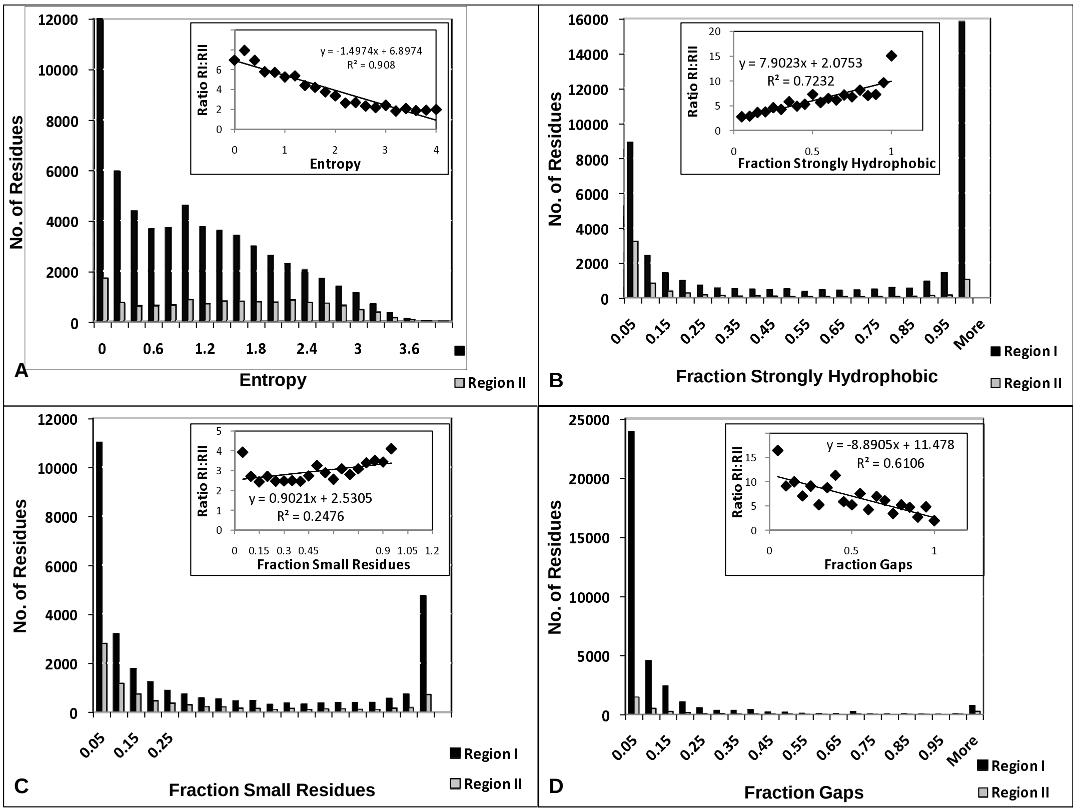


Figure 3.17: Bin frequency distributions of various fractional parameters for the learning set list. Frequency distribution plots of 73727 query residues from the set of 268 protein chains for A. Entropy B. Fraction strongly hydrophobic C. Fraction small residues D. Fraction gaps. Each distribution has been divided into its component major Region I and major Region II. All the inserts represent the ratio of major Region I to major Region II for the corresponding distributions.

change pattern at fractional value 0.50. Hence, 0.50 was taken as a threshold of the strongly hydrophobic fraction for the prediction of surface accessibility. Also, the relative ratio of the frequency of gaps and small residues change trend after fractional value 0.15, at which the ratio of the two is 1.00. Therefore, a conservative threshold of the small residue fraction was chosen as the fractional value 0.15 for the prediction of buried and surface accessible residues. The frequency of gaps becomes negligible above the fractional value of 0.50.

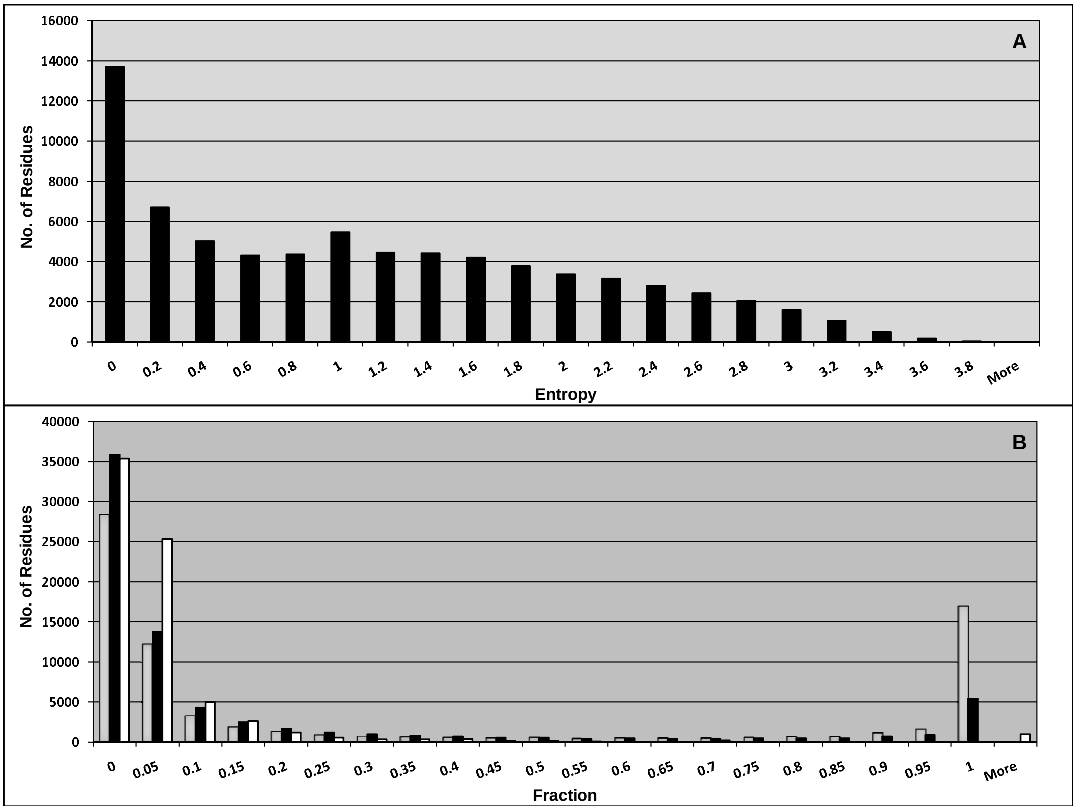


Figure 3.18: Overlay frequency distributions of learning set list for various fractional parameters. A. Frequency distribution of homology-based parameters for a set of 268 query protein chains. Frequency of a total of 73, 731 query residues, calculated over a total of 235, 138 alignments, with respect to each sequence entropy value. B. Frequency of query residues for fraction strongly hydrophobic (grey bars), fraction of residues that are small (black bars) and fraction of gaps (white bars) with respect to their corresponding values. All the fractional parameters are calculated over 235, 138 alignments corresponding to the 268 protein list.

3.3.3 Filter Parameter Based Analysis

The frequency distribution analysis of the various filter parameters, as chosen in subsection 3.3.2, have been evaluated here.

The frequency distributions of the two halves of the learning set list, that are bifurcated by the entropy filter threshold (≥ 0.15) have been shown in Figure 3.19. The density distributions of the residues clearly show a closer to normal frequency distribution for entropy ≥ 0.15 (Figure 3.19A) than for entropy < 0.15 (Figure 3.19B). A larger fraction of residues that satisfy the filter criteria, seem to fall in the lower packing density regions. The mean of the distribution lies between packing density $11 - 16$. The frequency distribution of the residues that do not satisfy the filter threshold condition, shown in Figure 3.19B, is a left skewed normal distribution. Here, the majority of residues appear to be concentrated towards higher packing densities. An overlay plot of the two distributions mentioned above is shown in Figure 3.19C. The cumulative frequency distribution, shown in Figure 3.19D, provides additional support for the filters. 28.80% of the aligned residues satisfy the filter criteria (Entropy ≥ 0.15) and 17.54% of the aligned residues do not satisfy the filter criteria (Entropy < 0.15) have density values less than 11 (surface accessible). Hence, an excess of 11.26% of the aligned residues is present on the surface in the lot that satisfy the filter criteria than the complementary lot (that fail the filter criteria). Hence, the entropy filter threshold of ≥ 0.15 should provide a more conservative first degree of filtering to the surface accessible residues.

The frequency distributions of small residues satisfying the filter threshold (FSR < 0.15) and its complementary lot (that satisfy the condition FSR ≥ 0.15), have been shown in Figure 3.20. The frequency distribution of residues satisfying the filter criteria, fraction small residues < 0.15, shown in Figure 3.20A, is a normal distribution with slight left skewedness. Its mean value seems to lie between packing density $15 - 16$. The frequency distribution of residues that are complementary to the filter

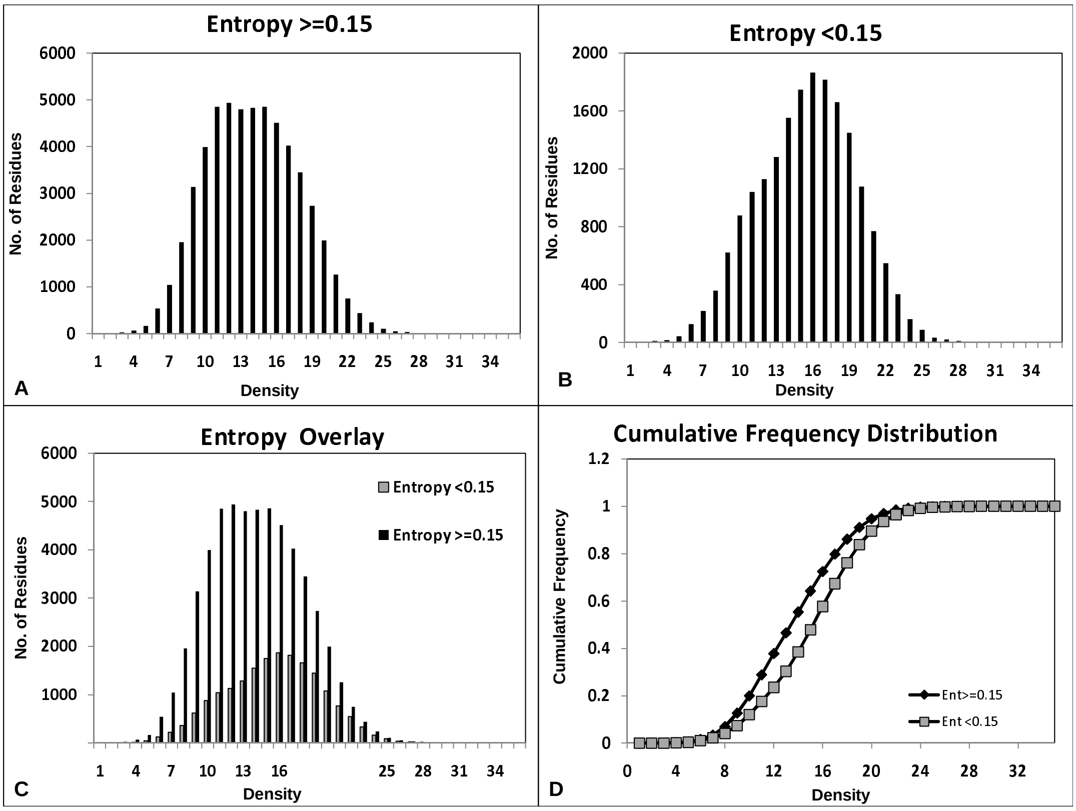


Figure 3.19: Threshold frequency distributions for entropy. Frequency distribution plots for the evaluation of entropy filter threshold. A. Frequency distribution plot of residues satisfying the condition- Entropy ≥ 0.15 (ordinate) at each packing density position (abscissa), B. Frequency distribution plot of residues satisfying the condition- Entropy < 0.15 (ordinate) at each packing density position (abscissa) C. An overlay plot of distributions plotted in Figure A and B (ordinate) at each density position (abscissa) and D. their cumulative frequency distribution (ordinate) at each packing density bin.

threshold (FSR ≥ 0.15), shown in Figure 3.20B, is a normal distribution that is slightly flattened at the top that ranges from packing density $10 - 18$. An overlay of these two plots is shown in Figure 3.20C. An overlay cumulative frequency distribution plot for the residues that satisfy the threshold criteria for fraction small residue filter and its complementary residue set (that fail the filter criteria) has been shown in Figure 3.20D. 32.88% of the aligned residues that satisfy the filter threshold criteria (FSR < 0.15), and 29.76% of residues that do not satisfy the filter threshold criteria (FSR ≥ 0.15) have packing densities lower than 11. Hence, the residues satisfying the filter threshold criteria have 3.12% more surface accessible residues than the complementary set.

The frequency distributions of strongly hydrophobic residues bifurcated by the threshold parameters have been presented in Figure 3.21. From the density distribution plot of residues satisfying the filter threshold criteria (Figure 3.21A), a right skewed normal distribution with majority of residues falling in low packing density regions (more surface accessible) is observed. The maxima of the curve is attained at packing density 11. The density distribution of residues with fraction strongly hydrophobic value ≥ 0.5 (Figure 3.21B) is a left skewed normal distribution. Here, the majority of the residues seem to be concentrated at higher packing densities. The maxima of this plot is observed at the packing density value of 16. An overlay plot of the above two mentioned distributions is shown in Figure 3.21C. An overlay cumulative frequency distribution plot has been shown in Figure 3.21D. The cumulative frequency curve that satisfies the filter criteria (FSHP < 0.5), indicates a presence of approximately 34.58% residues in likely surface accessible major Region II (packing density ≤ 11). The cumulative frequency distribution curve of residues that fail the filter criteria (FSHP ≥ 0.5), indicates a presence of approximately 12.53% residues in major Region II. Hence, the curve satisfying the filter threshold (FSHP < 0.5), has 22.05% more surface accessible residues than the complementary cumulative frequency distribution

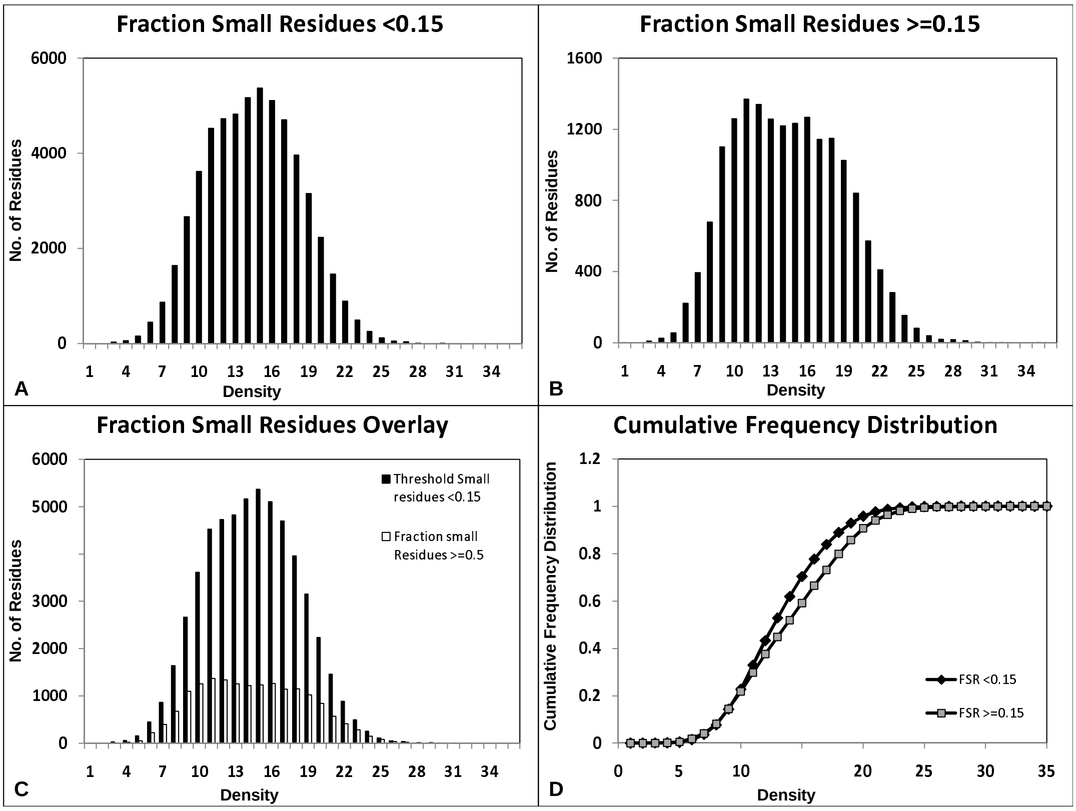


Figure 3.20: Threshold frequency distributions for Fraction Small Residues (FSR). Frequency distribution plots for the evaluation of fraction small residues (FSR) filter threshold. A. Frequency distribution plot of residues satisfying the condition- FSR $<$ 0.15 (ordinate) at each packing density position (abscissa), B. Frequency distribution plot of residues satisfying the condition- FSR $\geq$ 0.15 (ordinate) at each packing density position (abscissa) C. An overlay plot of distributions plotted in Figure A and B (ordinate) at each density position (abscissa) and D. their cumulative frequency distribution (ordinate) at each packing density bin (abscissa).

curve (≥ 0.5). This suggests that on the application of this filter threshold, the probability of a residue falling in major Region II (surface accessible) increases by at least two fold.

A comparison of the above mentioned cumulative frequency distribution plots, for the evaluation of three homology-based filter thresholds, shows filtering based on fraction strongly hydrophobic is expected to have highest contribution towards prediction accuracy. This is because the partitioning ratio between the distribution of residues that satisfy the filter criteria and that do not is maximum for fraction of strongly hydrophobic residues at packing density 11. Also, since all these filter thresholds have been selected conservatively and are biased towards the core of the protein, they are expected to have better prediction accuracy for the buried residues than for the surface accessible ones.

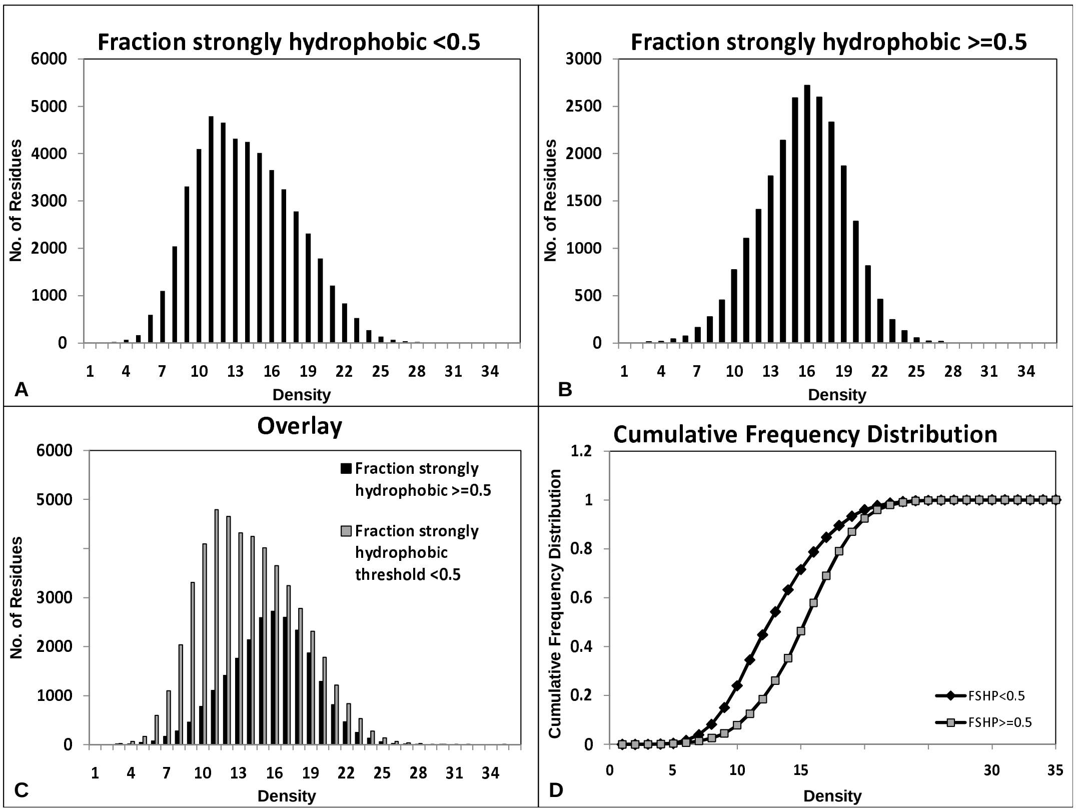


Figure 3.21: Threshold frequency distributions for Fraction Strongly Hydrophobic (FSHP). Frequency distribution plots for the evaluation of fraction strongly hydrophobic (FSHP) filter threshold. A. Frequency distribution plot of residues satisfying the condition- FSHP < 0.5 (ordinate) at each packing density position (abscissa), B. Frequency distribution plot of residues satisfying the condition- FSHP ≥ 0.5 (ordinate) at each packing density position (abscissa) C. An overlay plot of distributions plotted in Figure A and B (ordinate) at each density position (abscissa) and D. their cumulative frequency distribution (ordinate) at each packing density bin (abscissa).

Chapter 4

Discussion

In this chapter the results obtained from the aggregate plots and from the various frequency distributions have been evaluated in light of filter threshold selection for residue surface accessibility predictions. Additionally, an attempt to relate the aggregate trends of the various homology-based parameters with their physiological significance has also been presented here.

The structurally and sequencially diverse list of 268 protein chains, used in all the analysis in this work, is a representative subset of the entire collection of proteins. This is indicated by the nature of frequency distribution of residues at individual density positions. The histograms for the entire learning set of 268 proteins as well as the individual histograms for each of the three subsets of this list (monomeric, homodimeric and heterodimeric protein lists) have been found to have a Gaussian distribution. Inferring from the Central Limit Theorem, since,"the sum of a large number of independent random variables is distributed approximately normally,"[53] all the four protein learning sets used in this work are considered good representative of the proteins compiled in the databases. The filtering parameters implemented and evaluated, have been shown to be very effective in predicting core residues.[54]

4.1 Packing Density: A Measure of Compactness in a Protein

Although a coarse grain approach for the assessment of residue compactness in the native state of a protein, residue packing density used in this work provides a very good estimate of residue compactness. Better accuracy in the prediction of secondary structure by using three-dimensional coordinates of consecutive C_α atoms has been demonstrated.[55] Therefore, the use of C_α atoms in the calculation of packing density proves to be reasonable method. The increase in hydrophilicity value accompanied by a decrease in hydrophobicity value with increase in packing density suggests that residues with high packing density should be a part of the core and the residues with low packing density ought to be found at the surface of the protein[56,57] and possesses greater potential for mutability, flexibility[58] and ligand interactions. The aggregate trends of relative surface accessibility, derived from NACCESS,[31] further supports the notion that residue packing density provides an appropriate measure of relative surface accessibility.

As against the complex all atom simulation models that are time and computation intensive, coarse-grain approach provides more realistic and less complicated solutions where atom-depth resolution is not necessity.[59,60] However, in parsing relatively complex protein domains, this approach might prove to be more challenging.[59] Since the relevance of each single atom in the attainment of protein's native fold is limited, coarse-grain approaches like the one used in this work for the calculation of residue packing density can provide simple, but promising alternative for the theoretical expansion of knowledge of the complex bio-molecular system.[60] Our method does not treat voids and pockets explicitly and is not a measure of physical density. Although we assume that packing density takes into account both long and short range interactions, the contribution of local sequence information on surface accessibility prediction is noted to be insignificant.[57]

4.2 Aggregate Trend Analysis

4.2.1 Sequence Entropy: A Measure of Flexibility or Mutability

Sequence entropy is a measure of evolutionary mutability. Therefore, a correlation between sequence entropy and packing density[26] may be evident and also should be associated with the relative surface accessibility as discussed in section 4.1. The validity of aggregate entropy-inverse density trend of 40% sequence entropy used in this work was demonstrated by comparable aggregate trends of 6-point entropy with similar cutoffs and the HSSP derived entropy values. An inflexion point associated with the aggregate sequence entropy-inverse density trend indicates a point of demarcation between two major regions of the proteins associated with low and high packing densities. As discussed in section 4.1, these two regions- major Region I and major Region II, could possibly be characterized as buried (forming the core of the proteins) and surface accessible (forming the surface of the proteins) regions. Deviation in the aggregate trends from the learning set list was found to be maximum for the heterodimeric list. This can be attributed to the presence of a wide range of quaternary contacts in heterodimeric proteins which is absent in the other two types- monomeric and homodimeric proteins. Also, for heteropolymers, there are particular inconsistencies related to their lack of correlation involving short and long range interactions.[61] As there is a lack of consensus about the increase in RSA prediction accuracy by inclusion of evolutionary information from PSI-Blast,[62] the choice of Blastp appears to be a better alternative.

4.2.2 Fractional Parameters

From the overall aggregate trend analysis and frequency distributions of all the homology-based parameters for the learning set list of proteins, it is indicated that the core of the protein is packed with a combination of non-strongly hydrophobic residues and small residues. As we move out from this core towards the surface of the protein, within a particular density range ($22-14$), where the presence of strongly hydrophobic core (marked by a maxima in the aggregate trend of fraction of residues that are strongly hydrophobic) is expected, the frequency of aligned hydrophobic residues is maximum and the frequencies of aligned small residues and aligned non-strongly hydrophobic residues is minimum. Contrary to the finding that the association of nonpolar amino acids leads to the formation of protein's core that has dimensions similar to that of the randomly packed spheres at the percolation threshold,[21,63,64] the indications of strongly hydrophobic core has been observed here to occur at a critical distance from the most densely packed region of the protein. This region of strongly hydrophobic core formation is also found to involve a minimum occupancy or involvement of small and non-strongly hydrophobic residues. These results contradict the finding that the residues present at the topohydrophobic (that is strongly hydrophobic) positions are "very significantly more buried" than the residues that are non-topohydrophobic.[32]

The association of highest packing density values with a higher ratio of non-strongly hydrophobic (mostly polar) and small residues (glycine and alanine- nonpolar residues with smaller side chain) indicates that the core of the protein is a combination of polar residues whose hydrogen-bonding requirements are satisfied[50,58] and small non-polar residues whose relative fraction is constant[65] with a negligible involvement of the non-polar residues with large side groups (R). This notion is in agreement with the fact that in many cases, the protein core, that is the densely packed portion of the protein, comprises mostly of α - helix and β - sheets.[58] We must acknowledge here that packing and conformation are not tightly linked.[66] However, when present in the interior of the protein, the cooperative contribution of salt-bridges and regular hydrogen-bonding interactions, like those involving ionizable residues and polar residues, are known to have significant contributions to the protein stability.[67] Since "the efficient filling of space" has been suggested as an important factor in protein's unique structure determination,[68] and the fact that the protein's core more closely resembles a solid[21,63,64] with face-centered cubic (fcc) packing,[58] it is intuitive enough for the densest portion of the protein to be occupied by small residues.

Although still densely packed with high packing densities, the fraction of strongly hydrophobic residues form a cluster[69] at a distance from the densest packed core. This clustering involves negligible or minimal presence of small and polar residues. Since these residues are bigger in size, extremely efficient packing is not expected. This might be a reason that the packing in the protein's core is supposed to be two-thirds face centered cubic packing (fcc) and one third occupying random positions.[22]

If propensity of nonpolar residues alone is the driving force for protein folding, then it cannot be the principle factor in the determination of the protein's native fold.[66] This further raises the question whether protein's native fold was determined by residues that are not nonpolar, that is, more polar in character. A combined propensity of polar and non-polar residues is shown to provide a more logical protein folding initiation mechanism.[70] Better evolutionary conservation of residues involved in the formation of protein folding nucleus than all the buried residues[65] also indicates a possibility that the nucleation might involve a higher fraction of non-hydrophobic and small residues than strongly hydrophobic residues. Also, a correlation between packing efficiency, number of buried polar groups and protein stability has been noted.[71]

These results are consistent with the notion that the internal driving force of protein folding is a well balanced combination of two main factors- a) Propensity of the non-polar residues in the formation of the core[70,72] and b) Packing considerations associated with excluded volume.

The definition of the protein's core can be misleading if defined exclusively on the basis of sequence conservation,[73] but the residues with high packing density accompanied by high evolutionary conservation indicate involvement in either the protein folding nucleus or the stable core.[37] Core structure has also been identified by measures of hydrophobicity.[73]

The aggregate trend of the fraction of gaps, for the learning set list proteins, shows a steady increase with increase in inverse packing density, suggesting that the regions on the surface are more mutable than those lying in the protein core.

4.3 Filter Threshold Selection And Surface Accessibility Prediction

The various approaches to the residue solvent accessibility prediction include information theory, support vector machines, neural networks, nearest-neighbor methods, energy optimization and statistical analysis of amino acid composition.[56,57,74] Some of the drawbacks of these methods are not taking into account protein's structural information, using small datasets and unreliable sequence profile (by applying position specific scoring matrices (PSSM) in case of distant homologs).[74] These drawbacks can be easily overcome by the simple approach proposed in this thesis as it is a well balanced combination of homology-based parameters and packing density (structure-based concept). Optimization of parameters might pose some constraints on the prediction accuracy for example, in case of surface accessibility predictions by SABLE method[62,75] that use too many parameters.[76] It is due to this reason that such complicated prediction models might not prove to be better alternatives to the simpler ones. The predominant tendency of nonpolar residues to remain buried in the interior of the protein and the polar residues to exist on the surface[57] indicates the possibility of two state classification of protein residues in light of surface accessibility without proper consideration of protein secondary structure or their packing.[76]

A direct comparison of the various methods of surface accessibility prediction becomes challenging due to the difference in datasets, structure definition, threshold selection, number of categories and normalization schemes.[29,57,62] The support vector machine (SVM) approach is based on the neighboring residues as well as the physicochemical properties of amino acids in question.[77] For two-state prediction with a threshold of surface accessibility cutoff of 20%, the SVM and Neural Network (NN) methods had approximately 79% prediction accuracy while the performance of other methods like, the Baeysian Statistics (BS), Multiple linear regression (MLR) and Data tree (DT) achieved prediction accuracy of 71.2%, 71.6%, 71.5% respectively.[57] The homology-based modeling approaches are still on the cutting edge[54] with at least

82% prediction accuracy from sequence alignment methods and structural alignment methods.[57,78]

A common basis of the secondary structure prediction and classification of buried or surface accessible residue,[79] puts an upper limit ($70 - 80\%$ accuracy) on any such predictions made directly from the protein sequence.[62,75,80] The problem of buried/surface residue classification has some common basis with the classification of secondary structure elements.[79] Surface residue prediction, involves having to deal with the coupled nature of surface residue accessibility with quaternary structure.[81] The coupling of local secondary and higher orders of three dimensional structure poses intrinsic limitations on the prediction of residue surface accessibility. This is analogous to the difficulties of secondary structure prediction because of the coupling of secondary and tertiary structure in proteins.[54,81,82]

The characterization of two major regions (Region I and Region II), on the basis of packing density, as surface accessible and buried, suffers from noise and at the same time provides an inaccurate measure of surface accessibility predictions. Since the intrinsic propensity of each type of residue guides the attainment of local conformations to some extent,[76] the homology-based fractional parameters should serve as additional filters for surface accessibility predictions. Hence, a set of homology-based filters, based on entropy, fraction strongly hydrophobic, fraction small residues and fraction gaps could be employed for the surface accessibility.[83] Also, given the importance of residue hydrophobicity in the packing[26,84] of the core of the protein, it makes sense to include it as one of the key filters for homology modeling as a method to predict surface accessible residues. We assume that packing density, applied as a base parameter for predictive purposes, takes into account both the long and the short range interactions.[61,76] The inflection point (inverse density equal to 0.09) of the aggregate trend of sequence entropy versus inverse packing density plot was taken as a demarcation point between major Region I (packing density ≥ 11) and major Region

II (packing density > 11). A conservative approach was taken for the selection of best homology based filter thresholds.

The three homology-based filters - $Entropy \geq 0.15$, $FSR < 0.15$ and $FSHP < 0.5$, predict 11.26%, 3.12% and 22.05% more residues in the surface accessible major Region II than their counterpart. This further increases the binary prediction accuracy of surface accessible residues already characterized as present in the surface accessible major Region II. Filter based on fraction strongly hydrophobic is expected to remove most of the false positives because, highly skewed residue distribution obtained from the filtering parameters are known to increase the accuracy of prediction methods.[29] It is supported by the fact that in the two-state prediction of surface accessibility, hydrophobic residues are known to play key role.[85] The entropy based filters should provide a very conservative threshold for the surface accessibility predictions. The filter threshold chosen for fraction small residues is not expected to help much in improving the accuracy of the predictions but in combination with FSHP threshold, the FSR threshold can assist in refining the predictions and hence accuracy to some degree. Also, since larger fraction of residues are associated with major Region I, the prediction of buried residues is expected to give a better percentage of prediction accuracy. Since, two of the homology-based parameter filters are associated with nearly the same packing density region (major Region I), surface accessibility can directly be predicted from packing density. Also, it has been indicated that the residue solvent accessibility can directly be predicted from the three-dimensional coordinates of main and side chain atoms.[29]

The prediction accuracy of hydrophobic residues that are buried and polar residues that are surface accessible, is expected to be better than the rest of binary classified residues.[56] Although weak but positive influence of smaller protein chains on better prediction accuracies (smaller volume to surface area ratio associated with smaller protein chains), suggests optimization of this parameter in large learning datasets.[57]

Prediction accuracy also depends on the residue position, residue type and data size.[86] Therefore, to achieve better accuracy these parameters should also be scrutinized more carefully.

Where there are suggestions of preferences for certain higher order interactions in both the core and surface of proteins,[67] analysis of packing should then include more than pair-wise interactions.[87,88] Here we see the limits of averaging various homology-based values like sequence entropy and fraction of aligned non-strongly hydrophobic residues because, in part, they often do not fully account for the local environment of a particular residue.[89]

Chapter 5

Conclusions

Two major regions (major Region I and major Region II) have been consistently observed in the entropy-inverse density correlation plots for the structurally diverse set of 268 proteins, considered in this work. The major Region I, associated with high packing density (≥ 11) forms the core of the protein and hence all the residues associated with major Region I should be buried in the core of the protein. The major Region II, associated with low packing density values (< 11) are more exposed to the solvent and hence residues falling in Region II should be surface accessible. The regions with packing densities less than 4 and greater than 25 have been categorized as anomalous regions.

Attainment of maxima, by FSHP at the same packing density at which FSR and FNSHP attain a minima in the aggregate correlation plots, suggests that this general packing density region is associated with the formation of strongly hydrophobic core. The aggregate correlation plots also demonstrate that the highest packed region of protein is marked by the presence of small and non-strongly hydrophobic residues with a negligible presence of strongly hydrophobic residues. This implies that possibly the strongly hydrophobic core is not necessarily associated with the densest portion of the protein (where the fraction of small residues and non-strongly hydrophobic residues is minimum and the fraction of strongly hydrophobic residues is maximum).

The combined use of the three filter thresholds obtained from the homology-based parameters, namely, entropy (≥ 0.15), fraction small residues (< 0.15) and fraction strongly hydrophobic (< 0.5), should provide a conservative means of characterizing protein residues, as buried or surface accessible. Since the total number of residues in major Region I is more than the total number of residues in major Region II and also since these filter thresholds are chosen conservatively (biased towards major Region I), the prediction accuracy based on these filter thresholds is expected to be higher for the prediction of buried residues than for surface accessible ones.

Chapter 6

Future Studies

Based on the current work explained in this thesis, the issues that should be addressed in future studies have been summarized as below:

- The data set should be increased in such a way that it contains representative proteins with all the most likely fold types (800) present in the databases.[90]
- Inclusion of the loop residues in the prediction should also be checked for an increase in the quality of prediction.
- All the protein chains belonging to any given protein should be evaluated in a combined fashion.
- Sequence homology and other surface accessibility information for the prediction of quaternary contacts or vice versa should be explored (see Appendix Figure B.7).
- The effect of various thresholds on surface accessibility prediction accuracy should be evaluated.
- The tertiary and quarternary contacts should be evaluated for their dominant presence in the two major regions. All this information should be utilized for the development and understanding of residues found at the docking site of the proteins.
- The physical relevance of maxima obtained by the fraction of strongly hydrophobic residues in the inverse density plot should be explored in detail. Especially, the hypothesis that the nucleation of small and non-

strongly hydrophobic residues triggers the hydrophobic collapse or not should be evaluated in detail.

- Also, it should be checked if there exists any minimum number of small and non-strongly hydrophobic residues associated with the formation of a hydrophobic core.

- A characterization of protein types that are in agreement to the aggregate trends observed in this thesis and also that deviate from the aggregate behavior should be studied.

- These trend analyses should then be correlated to the type of protein secondary structures.

Bibliography

[1] Nambudripad, R.; Schmidt, F. T. *Kirk-Othmer Encyclopedia of Chemical Technology* **2005**, *0*, 1–21.

[2] Schueler-Furman, O.; Wang, C.; Bradley, P.; Misura, K.; Baker, D. *Science* **2005**, *310*, 638–642.

[3] Carugo, O. *Protein Eng.* **2000**, *13*, 607–609.

[4] Bonneau, R.; Baker, D. *Ann. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 173–189.

[5] Baker, D.; Sali, A. *Science* **2001**, *294*, 93–96.

[6] Dale, G. E.; Oefner, C.; DArcy, A. *J. Struct. Biol.* **2003**, *142*, 88–97.

[7] Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M. *Proc. Natl. Acad. Sci. USA.* **2007**, *104*, 9615–9620.

[8] Sanchez, R.; Pieper, U.; Melo, F.; Eswar, N.; Marti-Renom, M. A.; Madhusudhan, M. S.; Mirkovic, N.; Sali, A. *Progress* **2000**, 986–990.

[9] Fiser, A. *Expert Rev. Proteomics* **2004**, *1*, 97–110.

[10] Eswar, N.; Sali, A. *Protein structure modeling*; Springer Science + Business Media B. V., 2009.

[11] Kinch, L. N.; Wrable, J. O.; Krishna, S. S.; Majumdar, I.; Sadreyev, R. I.; Qi, Y.; Pei, J.; Cheng, H.; Grishin, N. V. *Proteins* **2003**, *53*, 395–409.

[12] Sippl, M. J. *Bioinformatics* **2008**, *24*, 872–873.

[13] Xu, Y.; Xu, D.; Liang, J. *Computational methods for protein structure prediction and modeling. Volume 2: Structure prediction*; Springer Science, 2007.

[14] Venclovas, C.; Zemla, A.; Fidelis, K.; Moult, J. *Proteins* **2003**, *53*, 585–595.

[15] Tanaka, S.; Scheraga, H. A. *Macromolecules* **1977**, *10*, 291–304.

[16] *Laboratory of Andrej Sali*, http://salilab.org/pdf/Eswar_NATOScience_2009.pdf. Accessed 2009.

[17] Sayano, K.; Shirakawa, T.; Kubota, Y.; Sarai, A. *RIKEN Review: Focused on Computational Science and Engineering* **1996**, *14*, 11–12.

[18] Dima, R. I.; Thirumalai, D. *Protein Sci.* **2006**, *15*, 258–268.

[19] *National Center for Biotechnology Information (NCBI), Protein Blast (BLASTP)*, http://www.ncbi.nlm.nih.gov/. Accessed 2008.

[20] Richards, F. M. *J. Mol. Biol.* **1974**, *82*, 1–14.

[21] Liang, J.; Dill, K. *Biophy. J.* **2001**, *81*, 751–766.

[22] Bagci, Z.; Jernigan, R. L.; Bahar, I. *J. Chem. Phys.* **2002**, *116*, 2269–2276.

[23] Gomes, M. A. F. *J. Phys. A: Math. Gen.* **1987**, *20*, L283–L284.

[24] Dima, R. I.; Thirumalai, D. *J. Phys. Chem. B* **2004**, *108*, 6564–6570.

[25] Liao, H. *Flexibility and sequence variability in proteins*; MS Thesis, San Jose State University, San Jose, 2004.

[26] Liao, H.; Yeh, W.; Chiang, D.; Jerniganf, R. L.; Lustig, B. *Protein Eng.* **2005**, *18*, 59–64.

[27] Koehl, P.; Levitt, M. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 1280–1285.

[28] Moret, M.; Zebende, G. F. *Phys. Rev. E* **2007**, *75*, 011920(1–4).

[29] Richardson, C. J.; Barlow, D. J. *Protein Eng.* **1999**, *12*, 1051–1054.

[30] Liu, Y. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1936–1941.

[31] Hubbard, S. J.; Thornton, J. M. *NACCESS, Computer Program, Department of Biochemistry and Molecular Biology, University College London.*, http://www.bioinf.manchester.ac.uk/naccess/, 1993.

[32] Poupon, A.; Mornon, J.-P. *FEBS Lett.* **1999**, *452*, 283–289.

[33] Chakrabarti, P.; Janin, J. *Proteins* **2002**, *47*, 334–343.

[34] Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J. *Proteins* **2003**, *53*, 708–719.

[35] Yeh, W. *Detailed analysis of protein sequence entropy, hydrophobicity, and flexibility*; MS Thesis, San Jose State University, San Jose, 2005.

[36] Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J. *J. Mol. Biol.* **2004**, *336*, 943–955.

[37] Guharoy, M.; Chakrabarti, P. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15447–15452.

[38] Wang, G.; Roland L. Dunbrack, J. *Bioinformatics* **2003**, *19*, 1589–1591.

[39] *The Protein Data Bank (PDB)*, http://www.pdb.org/pdb/home/home.do. Accessed 2008.

[40] Westbrook, J.; Yang, H.; Feng, Z.; Berman, H. M. *International Tables of Crystallography* **2006**, *G*, 539–543.

[41] Mirny, L. A.; Shakhnovich, E. I. *J. Mol. Biol.* **1999**, *291*, 177–196.

[42] Sander, C.; Schneider, R. *Nucleic Acids Res.* **1993**, *21*, 3105–3109.

[43] Dodge, C.; Schneider, R.; Sander, C. *Nucleic Acids Res.* **1998**, *26*, 313–315.

[44] Sander, C.; Schneider, R. *Proteins* **1991**, *9*, 56–68.

[45] Brandt, B. W.; Heringa, J.; Leunissen, J. A. M. *Nucleic Acids Res.* **2008**, *36*, W255–W259.

[46] Altschul, S. F.; Boguski, M. S.; Gish, W.; Wootton, J. C. *Nature Genetics* **1994**, *6*, 119–129.

[47] Hopp, T. P.; Woods, K. R. *Proc. Natl. Acad. Sci. USA* **1981**, *78*, 3824–3828.

[48] Engelman, D. M.; Steitz, T. A. *Cell* **1981**, *23*, 411–422.

[49] Sharp, K. A.; Nichollas, A.; Friedman, R.; Honig, B. *Biochemistry* **1991**, *30*, 9686–9697.

[50] Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534–552.

[51] Dill, K. A. *Biochemistry* **1985**, *24*, 1501–1509.

[52] Tanford, C. *J. Am. Chem. Soc.* **1962**, *84*, 4240–4247.

[53] Trivedi, K. S. *Probability and statistics with reliability, queuing and computer science applications.*; John Wiley and Sons, New York, 2001.

[54] Do, S.; Mishra, R. P.; Lakkaraju, H.; Dee, J.; Kantardjieff, K.; Lustig, B. Science, to be submitted for publication.

[55] Hosseini, S. R.; Sadeghi, M.; Pezeshk, H.; Eslahchi, C.; Habibi, M. *Comput. Biol. Chem.* **2008**, *32*, 406–411.

[56] Naderi-Manesh, H.; Sadeghi, M.; Arab, S.; Movahedi, A. A. M. *Proteins* **2001**, *42*, 452–459.

[57] Chen, H.; Zhou, H.-X.; Hu, X.; Yoo, I. Classification comparison of prediction of solvent accessibility from protein sequences. *Conferences in Research and Practice in Information Technology: 2nd Asia-Pacific Bioinformatics Conference (APBC2004)*, Dunedin, New Zealand, 2004.

[58] Rose, G. D. *Ann. Rev. Biophys. Biomol. Struct.* **1993**, *22*, 381–415.

[59] Zhang, Z.; Lu, L.; Noid1, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. *Biophy. J.* **2008**, *95*, 5073–5083.

[60] Clementi, C. *Curr. Opin. Struc. Biol.* **2007**, *17*, 1–6.

[61] Miyazawa, S.; Jernigan, R. L. *Proteins* **2003**, *50*, 35–43.

[62] Adamczak, R.; Porollo, A.; Meller, J. *Proteins* **2005**, *59*, 467–475.

[63] Moret, M. A.; Santana, M. C.; Nogueira, E.; Zebende, G. F. *Physica A* **2006**, *361*, 250–254.

[64] Moret, M. A.; Zebende, G. F. *Phy. Rev. E.* **2007**, *75*, 011920(1–4).

[65] Shakhnovich, E.; Abkevich, V.; Ptitsyn, O. *Nature* **1996**, *379*, 96–98.

[66] Behe, M. J.; Lattman, E. E.; Rose, G. D. *Proc. Natl. Acad. Sci. USA.* **1991**, *88*, 4195–4199.

[67] Li, X.; Liang, J. *Proteins* **2005**, *60*, 46–65.

[68] Richards, F. M. *CMLS-Cell. Mol. Life S.* **1997**, *53*, 790–802.

[69] Chowriappa, P.; Dua, S.; Kanno, J.; Thompson, H. W. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2010**, *6*, 639–651.

[70] Dyson, H. J.; Wright, P. E.; Scheraga, H. A. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 1305713061.

[71] Demarest, S. J.; Zhou, S.-Q.; Robblee, J.; Fairman, R.; Chu, B.; Raleigh, D. P. *Biochemistry-US* **2001**, *40*, 2138–2147.

[72] Daggett, V.; Fersht, A. R. *Trends Biochem. Sci.* **2003**, *28*, 18–25.

[73] Gerstein, M.; Altman, R. B. *J. Mol. Biol.* **1995**, *251*, 161–175.

[74] Bondugula, R.; Xu, D. *Computational Systems Bioinformatics* **2008**, *7*, 195–202.

[75] Adamczak, R.; Porollo, A.; Meller, J. *Proteins* **2004**, *56*, 753–767.

[76] Go, N. *Ann. Rev. Biophys. Bioeng.* **1983**, *12*, 183–210.

[77] Ogul, H.; Mumcuoglu, E. U. *LNAI* **2006**, *3949*, 141–148.

[78] Rost, B.; Sander, C. *Proteins* **1994**, *20*, 216–226.

[79] Nguyen, M. N.; Rajapakse, J. C. *Int. J. Data Min. Bioin.* **2007**, *1*, 248–269.

[80] Frishman, D.; Argos, P. *Proteins* **1997**, *27*, 329–335.

[81] Rost, B. *J. Struct. Biol.* **2001**, *134*, 204–218.

[82] Kihara, D. *Protein Sci.* **2005**, *14*, 1955–1963.

[83] Do, S.; Lakkaraju, H.; Potluri, S.; Pham, K.; Kantardjieff, K.; Lustig, B. *52nd Biophysical Society Meeting Abstracts. Biophys. J. Supplement, Abstract* **2008**, *94*, 3280–Pos.

[84] Jernigan, R. L.; Kloczkowski, A. *Methods Mol. Biol.* **2007**, *350*, 251–276.

[85] Yan, C.; Wu, F.; Jernigan, R. L.; Dobbs, D.; Honavar, V. *The Protein Journal* **2008**, *27*, 59–70.

[86] Ahmad, S.; Gromiha, M. M.; Sarai, A. *Proteins* **2003**, *50*, 629–635.

[87] Munson, P. J.; Singh, R. K. *Protein Sci.* **1997**, *6*, 1467–1481.

[88] Betancourt, M. R.; Thirumalai, D. *Protein Sci.* **1999**, *8*, 361–365.

[89] Bandyopadhyay, D.; Mehler, E. L. *Proteins* **2008**, *72*, 646–659.

[90] *Wellcome Trust Sanger Institute.*, ftp://ftp.sanger.ac.uk/pub4/theses/yeats/. Accessed 2009.

Appendices

Appendix A

Program Listings

```
#---------------------------------------------------------------------------------------------------
# Author: Radhika Pallavi Mishra
# Date: August 14, 2008.
# Purpose: Downloads mmCIF files from RCSB Protein Database
# File: ftp-script-1.pl
#---------------------------------------------------------------------------------------------------
#!/usr/bin/perl -w
use Net::FTP;
 sub doFTP
 {
   my ($line1) = @_ ;
   print "arg recieved, $line1 \n";
   chomp($line1);
   $line1 = lc($line1);
   $subdir = substr($line1,1,2);
   $destDir = "/pub/pdb/data/structures/divided/mmCIF/".$subdir;
   print "$destDir\n";
   $ftp->cwd($destDir);
   $ftp->binary();
   $filetoftp = $line1."\.cif\.Z";
   print "$filetoftp\n";
   $ftp->get($filetoftp,$filetoftp);}
 open(FHNDL,"SeqIDlearningset286.txt");
 $line = <FHNDL>;
 $ftp = Net::FTP->new("ftp.rcsb.org");
 $ftp->login("anonymous",'-anonymous@');
 while(!eof(FHNDL)){
   doFTP($line);
   $line = <FHNDL>; }
 doFTP($line);
$ftp->quit;
close(FHNDL);
end;

#------------------------------------------------------------------------------------------------
#------------------------------------------------------------------------------------------------

#------------------------------------------------------------------------------------------------
```

```
# Author: Radhika Pallavi Mishra

# Date: August 20, 2008

# Purpose: Takes mmcif file in ASCII text format and calculates density. Also identifies missing

# residues in ATOM section of the PDB files.

# Adapted from: pdb2denMOD-2 written by William Yeh

# File: cif2den.pl

#---------------------------------------------------------------------------------------------------


# User Specified Variables to Control Analysis and Output

@TabValues = (6, 7, 8, 9, 10, 11, 12); # For each Value, calc’s #dist <= Value

                                       # Must be increasing in value.

@TabPrint  = (0, 1, 2, 3, 4, 5, 6); # Defines which @TabValues printed

# Initialize Amino Acid 3-letter to 1-letter associative list

%AADictionary = (

  ’GLY’ => ’G’,  ’ALA’ => ’A’,  ’VAL’ => ’V’,  ’LEU’ => ’L’,

  ’ILE’ => ’I’,  ’MET’ => ’M’,  ’PRO’ => ’P’,  ’PHE’ => ’F’,

  ’TRP’ => ’W’,  ’SER’ => ’S’,  ’THR’ => ’T’,  ’ASN’ => ’N’,

  ’GLN’ => ’Q’,  ’TYR’ => ’Y’,  ’CYS’ => ’C’,  ’LYS’ => ’K’,

  ’ARG’ => ’R’,  ’HIS’ => ’H’,  ’ASP’ => ’D’,  ’GLU’ => ’E’ );


if($#ARGV < 1){

print "USAGE : perl cif2den.pl <Chain Name i.e. A,B etc..> <Name of .cif file> \n";

exit;}

$p_directory = "CIF_files";

# Initialize all variables

$Name      = "";   # PDB name (extracted from HEADER line), in lower case

$Line      = "";

@AtomLine  = ();   # Temp var for current ATOM line as list

@AASeqRes  = ();   # 3-letter form from SEQRES, for checking

@AA        = ();   # 3-letter form from ATOM statements

@AA1       = ();   # 1-letter form (translated)

@Tag       = ();   # Unused for now

@Distance  = ();   # Calc’d distances

$AAref     = ’’;

#$AArefI    = ’A’;

$Indexref  = 0;

$Tagref    = ’’;

$AACount   = 0;    # Number of aa’s

$Xref      = 0;

$Yref      = 0;

$Zref      = 0;

$i         = 0;

$j         = 0;

$k         = 0;
```

```
@X         = ();
@Y         = ();
@Z         = ();
@XVect     = ();
@YVect     = ();
@ZVect     = (); # Temp var for storing vector of aa being compared
@TabCount  = ();   # TabCount tracks no. < each @TabValues
@TabDensity= ();   # Calc density from TabCount & TabValues
@PrintLine = ();   # Temp var holds line for printing
$Count     = 0;
open ( IN  ,"$p_directory/$ARGV[1]") or die "Cannot open input files for read"."\n";
$i = 0;                                    # Indices for AA’s
while( <IN> ) {
$Line = $_ ; chomp($Line);               # Save current line
# =====  Extract Amino Acid seq from SEQRES statements  =====
if ($Line =~ /_pdbx_poly_seq_scheme\.pdb_ins_code/) {
                  while((!eof(IN))&&($Line !~ /loop_/)) {
  $Line = <IN> ;
                    chomp($Line);
                      @LineArray = split(/ +/,$Line);
                   if($LineArray[9] eq $ARGV[0]) {
push(@AASeqRes, $AADictionary{$LineArray[3]});
                            push(@FASTAPOS, $LineArray[4]);
                            push(@PDBPOSSEQRES, $LineArray[6]);}}}
   # =====  Extract alpha-C (x,y,z) from ATOM statements  =====
   if ($Line =~ /_atom_site\.pdbx_PDB_model_num/){
 while((!eof(IN))&&($Line !~ /\#/)) {
                   $Line = <IN> ;
                    chomp($Line);
if($Line =~ /^ATOM +[\d]+ +[A-Z]+ +CA +/ ) { # Find alpha-Carbon ATOM lines
     @AtomLine = split(/ +/, $Line);
     ($AAref, $Xref, $Yref, $Zref) = @AtomLine[5,10,11,12];
$NextLine = <IN>;
@NextLineArray = split(/ +/,$NextLine);
                              @NewAtomline = (@AtomLine,@NextLineArray);  # Radhika 09/10/08
                              $AAREFI = @NewAtomline[22];  # Radhika 09/10/08
                              $pdbpos = $NewAtomline[20];
if ( $AADictionary{$AAref} ne ’’ && ($AAREFI eq $ARGV[0])){ # ONLY EXTRACT A PARTICULAR CHAIN
   push( @AA , $AAref);
   push( @AA1, $AADictionary{$AAref});
push( @X  , $Xref);
   push( @Y  , $Yref);
   push( @Z  , $Zref);
   push( @PDBPosArray, $pdbpos);
print " AAREF is $AAref, AAREFI is $AAREFI, pdbpos is $pdbpos and x,y,z are $Xref, $Yref and $Zref \n";}}}}
# ===== Extract PDB name from HEADER line
if ($Line =˜ /ˆdata_/) {
chomp($Line);
     @AtomLine = split( /_/, $Line);
     $Name = $AtomLine[1];
     $Name = lc($Name);
                        print "Name is $Name \n";};};
$AACount = 1 + $#AA;
@TabValues = sort { $a <=> $b } @TabValues; # Make sure it’s ascending
# Output Filename...
open ( OUT ,’>’."$Name".$ARGV[0].".den") or die "Cannot open out_pdbden.txt for write.\n";
# =====  Calculate distances and tabulate  =====
# Note that $i is the aa location, and $j is used to scan to build vects.
for ($i = 0; $i < $AACount; $i++) {
for ($j = 0; $j < $AACount; $j++) {
@Distance[$j] = sqrt( (($X[$j] - $X[$i])**2)
     +(($Y[$j] - $Y[$i])**2)
     +(($Z[$j] - $Z[$i])**2) ); };
   # =====  Sort and tabulate according to distance
   for ($j = 0; $j <= $#TabValues; $j++) {
     $Count = 0;
     for ($k = 0; $k <= $#Distance; $k++) {
       if ($Distance[$k] <= $TabValues[$j]) {
$Count++ };}
     $TabCount[$j] = $Count;
     $TabDensity[$j] = 1000 * $Count / ((4.0/3.0)*3.14159
* ($TabValues[$j]**3)); # Compute density
}
# ===== Store density values in a hash corresponding to their PDB Position =====
$valueForHash = "";
foreach $i (@TabPrint) {                             # Output count C()
                        $valueForHash = $valueForHash."_".$TabCount[$i];};
                  $posDenHash{$PDBPosArray[$i]} = $valueForHash ;};
$numSeqResEntries = $#AASeqRes + 1;
                print "Number of residues in Sequence = $numSeqResEntries \n" ;
print OUT "Number of residues in Sequence = $numSeqResEntries \n" ;
for($k=0; $k < $numSeqResEntries; $k++){
if($PDBPOSSEQRES[$k] eq "\?"){
$PrintLine = "D "
         .$Name
."_".sprintf("%03d", $FASTAPOS[$k])
```

```
."_".sprintf("%1s " , $AASeqRes[$k]);
     foreach $l (@TabPrint) {                             # Output count C() = NA
$PrintLine = $PrintLine
."C(". $TabValues[$l] .")= "
."NA" ." " ;};
                          printf OUT  $PrintLine ." \?\n";}
 else{
   $PrintLine = "D "
         .$Name
."_".sprintf("%03d", $FASTAPOS[$k])
."_".sprintf("%1s " , $AASeqRes[$k]);
$density = $posDenHash{$PDBPOSSEQRES[$k]};
                               print "density is $density \n";
                                @densityArray = split(/_/, $density);
                                $m = 1;
foreach $n (@TabPrint) {                             # Output count C()
$PrintLine = $PrintLine
."C(". $TabValues[$n] .")= "
.sprintf("% 3d", $densityArray[$m]) ." " ;
                                         $m++;};
printf OUT  $PrintLine ." $PDBPOSSEQRES[$k]\n";}}
close(IN);
printf OUT "\n\n";
close(OUT);
```

```
-------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------
#Author: Radhika Pallavi Mishra
# Date: August 30, 2008
# Purpose: Calculates packing density for protein list with specified chains
# File: Chainselectivecif2den.pl
-------------------------------------------------------------------------------------------------
```

```
#!/usr/bin/perl -w
open(FHNDL,"SeqIDlearningset268.txt") or die "Cannot open filename " ;
$lines = <FHNDL>;
while(!eof(FHNDL)){
  chomp($lines);
  $identifier = substr($lines,4,1);
$filename = substr($lines,0,4);
$filename= $filename."\.cif";
#print " filename is $filename\n";
      #print "identifier is $identifier and line is $lines\n";
```

```
    system("perl cif2den.pl $identifier $filename" );
   $lines = <FHNDL>;}
close(FHNDL);
end;
```

```
-----------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------

-----------------------------------------------------------------------------------------------
#File: bst2entMOD2.pl
#! /usr/perl
# File: bst2ent   v4 9/12/2002 D.Chiang
# "Blast to Entropy Calculation"
#
# Usage: 'perl bst2ent.pl <PDB-name> <infile> <outfile> (<logfile>)'
#      Eg 'perl bst2ent.pl 1agm 1agmibst.txt 1agmoent.txt 1agmobst2entlog.txt'
#       generates '1agmoent' from processing '1agmibst.txt'.
#      The PDB name is used to tag all residues, which is then used to match the
#       PDB post-processing done by Perl script 'pdb2den.pl'. Therefore, the
#       protein-name MUST BE IDENTICAL to that in the corresponding PDB file.
#      Also generates logfile (default 'bst2entlog.txt') = a superset
#       of outfile for data verification purposes.
#      BLAST output file <infile> must be saved in Text format.
#
# Action: Takes BLAST output file in ASCII test format
#
#   1. Extracts all Query and Subject sequence pairs
#   2. Compacts the Query/Subject sequences back to length(Query)
#        by omitting all insertions in the Subject sequence.
#        (Deletions in the Subject seq are kept.)
#   3. Puts the Query and all Subject seqs in output matrix.
#   4. Calculates entropy values for "qualified" database sequences (ie such
#        as those with Identity% scores higher than $IdenPercentMin) and
#        "qualified" positions (ie there are a sufficient number of homologs
#        with non-deletions in that position, as specified in $HomologsMin)
#   5. Qualified sequences are specified in upper case residue codes, while
#        unqualified sequences use lower case codes.
#      Unqualified positions are flagged with entropy value output of "-1".
#
# Detail Notes:
#   1. Relies on first line chars being " Score =" to flag beginning
#        section of Query and Subject sequences. ScoreBits, Expect, IdenPercent,
#        and PositivePercent values are extracted. These are used to flag
#        whether sequence is counted in entropy calculations.
#   2. Concatenates seq sections until 2 blank lines encountered.
#   3. Format of *.out file:
#        1st line is original Query (extracted from 1st match pair)
#        Subsequent lines are Subject sequences modified by deleting
#        insertions and filling out both prefix and suffix to have
#        same length as original Query.
#   4. Note that BLAST can substitute ’X’ (proteins) or ’N’ (nucleotides) into
#        the Query sequence to filter out "low complexity" regions. These
#        residues are kept as X or N in the entropy calculation. However, they
#        can be post-processed when correlated with the PDB information using
#        the residue position number. They are converted to lower case
#        in the output (trick to help merging with pdb2den.pl output,
#        since lowercase sorts after all upper case).
#
# Revision History
#  v1.0 6/27/02  Initial version.
#  v1.1 6/28/02  Minor addition of ’;’ to output file.
#  v2.0 6/28/02  Change output format to transposed form.
#  v3   7/18/02  Change name to bst2ent.pl (from bl2seq.pl).
#                 - Use " Score =" (not "Query") to flag sequence sections.
#                 - Add entropy calculations.
#                 - Clean up misc code
#  v4   9/12/02  Extracted all ’-’ from Query sequence reported from 1st
#                 match in BLAST, to take care of case when the 1st match
#                 includes insertions (ie the query itself is not found).
#                Also reduced Identity% and #HomologMin parameters
#
#08/05/08 Modified by Radhika Pallavi Mishra to include the chain name
#and bitcutoff set to 0.
#
# Specify User Parameters
$IdenPercentMin = 0.10;       # Min value of Identity% score for qualified seq
# (IdenPercentMin is another approach of BLAST results cut-off. Not used for this study)
$HomologMin     = 1;          # Min value of non-deleted homologs for entropy calc
# (HomologMin is needed for entropy calc, so an error won’t occur when dividing by 0)
$ScoreMin       = 100;   # Min value of match score
$BitCutOff      = 0;
$p_directory = "nblast_all";
opendir (DIRECTORY, $p_directory);
while (defined($p_filename = readdir(DIRECTORY))) {
  if ($p_filename != "." || $p_filename != "..") {
      # Initialize all variables
      $PDBName        = "";          # Passed param; used to tag residues in output only
```

```
   $Line           = "";
   @LineSplit      = ();
   $ScoreBits      = 0;
   $Expect         = 0;
   $IdenPercent    = 0;
   $PositivePercent= 0;
   $QuerySeq       = "";         # Original Query Seq (from 1st pair)
   $QuerySeqTemp   = "";         # Query Seq in pair for manipulation
   $SubjectSeqTemp = "";         # Subject Seq in pair for manipulation
   $QueryOffset    = 0;          # Initial seq offset (always 1 for 1st seq)
   $LengthDiff     = 0;          # Temp vars ...
   $QueryOffDummy  = 0;
   $Dummy          = "";
   $Dummy2         = "";
   @SeqList        = ();
   $LineOut        = "";
   %EntropyCount   = ();
   $EntropyCountTot= 0;
   $Prob           = 0;
   $Entropy        = 0;
   # Setup files for writing and reading
open ( IN  ,"$p_directory/$p_filename") or die "Cannot open input files for read"."\n";
   $PDBName = substr($p_filename,0,5);         # Get PDB protein name as 1st parameter
  open ( OUT ,’>’.$PDBName."_".$BitCutOff.".ent") or die "Cannot open output file for write.\n";
open ( OUTD ,’>’.$PDBName."_".$BitCutOff.".dbg") or die "Cannot open debug file for write.\n";
   while( <IN> ) {
     $Line = $_ ; chomp($Line);                 # Save current line
     # ----------Extracting BitScore, Expected Value, Identity, Positives----------
     while ( $Line =˜ /ˆ Score =/ ) {           # Find next set of Query/Sbjct
         $QueryOffset    = 0 ;                   # Reset values
         $QuerySeqTemp   = "";
         $SubjectSeqTemp = "";
         @LineSplit = split(/ +/, $Line);
         $ScoreBits   = $LineSplit[3];
         $Expect      = $LineSplit[8];           # Extract Expect value
         if ($Expect =˜ /ˆe/) { $Expect = "1".$Expect }; # If format "e-xxx" add ’1’ prefix
         $Line = <IN>; chomp($Line);             # Save next line
         @LineSplit = split(/ +/, $Line);
         $IdenPercent = $LineSplit[4];           # Extract Identity%, stripping
         $IdenPercent =˜ s/[(),%]//g;            # strip out of "(xxx%)," format
         $IdenPercent = $IdenPercent / 100.0;
         $PositPercent= $LineSplit[8];           # Same with Positive%
         $PositPercent =˜ s/[(),%]//g;
```

```
$PositPercent = $PositPercent / 100.0;
$Line = <IN>;                           # Skip next line (should be blank)
$Line = <IN>;
# ----------Extracting Query Sequences, Subject Sequences, Query Offsets----------
while ( $Line =˜ /ˆQuery  / ) {         # Find 1st Query line in set
  ($Dummy, $QueryOffDummy, $Line, $Dummy2) = split(/ +/, $Line);  # Separate into fields
  if ($QueryOffset == 0) {             # Keep 1st $QueryOffset
    $QueryOffset = $QueryOffDummy };
  $QuerySeqTemp = $QuerySeqTemp.$Line;  # Combine running seq
  $Line = <IN>;                         # Throw away 2nd line
  $Line = <IN>; chomp($Line);           # Save 3rd = Sbjct line
  $Line =˜ s/ˆSbjct  \d+ +//;           # Strip seq prefix
  $Line =˜ s/ *\d+.*$//;                 # Strip suffix
  $SubjectSeqTemp = $SubjectSeqTemp.$Line; # Combine running seq
  $Line = <IN>;                         # Throw away 2nd line
  $Line = <IN>; chomp($Line);           # Next line (another Query?)
};
if ( $QuerySeq eq "") {                 # Very 1st Query is saved
  $QuerySeq = $QuerySeqTemp;
  $QuerySeq =˜ s/X/x/g;                 # However, convert special X,N chars to lower case
  # $QuerySeq =˜ s/N/n/g;               # (Should be removed, N is used for nucleotides only)
  $QuerySeq =˜ s/-//g;                  # Also, extract insertations (’-’)
  $ScoreMin = $ScoreBits*$BitCutOff/100;
  printf OUTD "# === Original Sequence (from 1st match)\n";
  printf OUTD $QuerySeq . "\n\n";
  printf OUTD "MaxScore = ";
  printf OUTD $ScoreBits;
  printf OUTD ", MinScore = ";
  printf OUTD $ScoreMin . "\n\n";
  push(@SeqList, $QuerySeq."\n");      # Storing Sequences for Entropy calculations
};
# =====  WRITE DEBUG FILE  =====
printf OUTD "# === NEW MATCH PAIR   Offset = " . $QueryOffset . "\n";
printf OUTD $QuerySeqTemp   . "\n";     # Write out complete seq
printf OUTD $SubjectSeqTemp . "\n\n";
# =====  PROCESS SEQUENCES =====
# Fill out prefix offset
$QuerySeqTemp = substr($QuerySeq, 0, -1+$QueryOffset) . $QuerySeqTemp;
$SubjectSeqTemp = ("-" x (-1+$QueryOffset)) . $SubjectSeqTemp;
printf OUTD "# === Matched pair with prefix & suffix filled\n";
printf OUTD $QuerySeqTemp   . "\n";     # Write out complete seq
printf OUTD $SubjectSeqTemp . "\n\n";
# Find & delete insertions
```

```
      for ($i = -1+length($QuerySeqTemp); $i >= 0; $i += -1) {
        if ( substr($QuerySeqTemp, $i, 1) eq "-" ) {
          substr($QuerySeqTemp, $i, 1) = "";
          substr($SubjectSeqTemp, $i, 1) = "";};};
      # Fill out suffix if necessary
      $LengthDiff = length($QuerySeq) - length($SubjectSeqTemp);
      if ( $LengthDiff > 0 ) {
        $QuerySeqTemp   = $QuerySeqTemp . substr($QuerySeq, - $Length);
        $SubjectSeqTemp = $SubjectSeqTemp . ("-" x $LengthDiff);};
      # =====  QUALIFY THIS SEQUENCE  =====
      # if ($IdenPercent < $IdenPercentMin) {   # If not-qualified, flag as lower case
      if ($ScoreBits < $ScoreMin) {   # If not-qualified, flag as lower case
        $SubjectSeqTemp = lc($SubjectSeqTemp);};
      # =====  WRITE OUT SEQUENCE  =====
      printf OUTD "# === Matched pair with insertions omitted\n";
      printf OUTD $QuerySeqTemp   . "\n";     # Write out complete seq
      printf OUTD $SubjectSeqTemp . "\n\n";
      # ===== REMOVE UNQUALIFIED SEQS FROM THE FINAL ENTROPY FILE OUTPUT =====
      if ($ScoreBits >= $ScoreMin) {
        push(@SeqList, $SubjectSeqTemp."\n");};};};
close(IN);
for ($i = 0; $i <= -2+length( @SeqList[0] ); $i += 1) { # Why is the -2 value there?
   $LineOut = "";
   for ($j = 1; $j <= $#SeqList; $j += 1) {
      $LineOut = $LineOut . substr( @SeqList[$j], $i, 1); # Cycling thru seq j, at pos i
   };
   # =====  COMPUTE ENTROPY  =====
   $Line = $LineOut;
   $Line =˜ s/[ˆA-Z]//g;                # Delete anything not capital (IMPORTANT!!!!)
   @LineSplit = split("", $Line);
   %EntropyCount    = ();
   $EntropyCountTot = 0;
   foreach $i (@LineSplit) {
      $EntropyCount{$i}++ ;
      $EntropyCountTot++  ; };
   @AllEntValues = sort(values(%EntropyCount));
   if ($EntropyCountTot >= $HomologMin) {
      $Entropy = 0;
      for ($j = 0; $j <= $#AllEntValues; $j++) {
         $Prob = $AllEntValues[$j] / $EntropyCountTot ;
         $Entropy = $Entropy - ($Prob * ( log($Prob)/log(2) ));}
      # Debug Code
      printf OUTD "@"
```

```
               ." Entropy= "        .$Entropy
               ." EntropyCount= "   .join(" ",%EntropyCount)
               ." EntropyCountTot= ".$EntropyCountTot
               ." AllEntValues= "   .join(" ",@AllEntValues)
               ."\n";}
         else {
            $Entropy = -1;                    # Flag as error -- too few homologs
         };
         $LineOut = "E= ".sprintf("% .3f",$Entropy)
                  ." A= ".$LineOut;
         # =====  CREATE OUTPUT LINE PREFIX  =====
         $LineHeader = "D "                        # Start line format eg "D 1agm_001_A"
                        .$PDBName."_"
                        .sprintf("%03d",1+$i)
                        ."_". substr(@SeqList[0], $i, 1);
         printf OUT $LineHeader ." ". $LineOut . "\n";};
      close(OUTD);
      close(OUT);};};
-----------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------
#File: Radhika-6pointentropy.pl
#Date : June 17, 2008
#! /usr/perl
# File: bst2ent.pl   v4 9/12/2002 D.Chiang
# "Blast to Entropy Calculation"
# Usage: ’perl bst2ent.pl <PDB-name> <infile> <outfile> (<logfile>)’
#      Eg ’perl bst2ent.pl 1agm 1agmibst.txt 1agmoent.txt 1agmobst2entlog.txt’
#       generates ’1agmoent’ from processing ’1agmibst.txt’.
#      The PDB name is used to tag all residues, which is then used to match the
#       PDB post-processing done by Perl script ’pdb2den.pl’. Therefore, the
#       protein-name MUST BE IDENTICAL to that in the corresponding PDB file.
#      Also generates logfile (default ’bst2entlog.txt’) = a superset
#       of outfile for data verification purposes.
#      BLAST output file <infile> must be saved in Text format.
#
# Action: Takes BLAST output file in ASCII test format
#
#   1. Extracts all Query and Subject sequence pairs
#   2. Compacts the Query/Subject sequences back to length(Query)
#        by omitting all insertions in the Subject sequence.
#        (Deletions in the Subject seq are kept.)
#   3. Puts the Query and all Subject seqs in output matrix.
#   4. Calculates entropy values for "qualified" database sequences (ie such
#        as those with Identity% scores higher than $IdenPercentMin) and
#        "qualified" positions (ie there are a sufficient number of homologs
#        with non-deletions in that position, as specified in $HomologsMin)
#   5. Qualified sequences are specified in upper case residue codes, while
#        unqualified sequences use lower case codes.
#      Unqualified positions are flagged with entropy value output of "-1".
#
# Detail Notes:
#   1. Relies on first line chars being " Score =" to flag beginning
#        section of Query and Subject sequences. ScoreBits, Expect, IdenPercent,
#        and PositivePercent values are extracted. These are used to flag
#        whether sequence is counted in entropy calculations.
#   2. Concatenates seq sections until 2 blank lines encountered.
#   3. Format of *.out file:
#        1st line is original Query (extracted from 1st match pair)
#        Subsequent lines are Subject sequences modified by deleting
#        insertions and filling out both prefix and suffix to have
#        same length as original Query.
#   4. Note that BLAST can substitute ’X’ (proteins) or ’N’ (nucleotides) into
#        the Query sequence to filter out "low complexity" regions. These
#        residues are kept as X or N in the entropy calculation. However, they
#        can be post-processed when correlated with the PDB information using
#        the residue position number. They are converted to lower case
#        in the output (trick to help merging with pdb2den.pl output,
#        since lowercase sorts after all upper case).
#
# Revision History
#  v1.0 6/27/02  Initial version.
#  v1.1 6/28/02  Minor addition of ’;’ to output file.
#  v2.0 6/28/02  Change output format to transposed form.
#  v3   7/18/02  Change name to bst2ent.pl (from bl2seq.pl).
#                 - Use " Score =" (not "Query") to flag sequence sections.
#                 - Add entropy calculations.
#                 - Clean up misc code
#  v4   9/12/02  Extracted all ’-’ from Query sequence reported from 1st
#                 match in BLAST, to take care of case when the 1st match
#                 includes insertions (ie the query itself is not found).
#                Also reduced Identity% and #HomologMin parameters
#
# Specify User Parameters
$IdenPercentMin = 0.10;       # Min value of Identity% score for qualified seq
# (IdenPercentMin is another approach of BLAST results cut-off. Not used for this study)
```

```
$HomologMin     = 1;          # Min value of non-deleted homologs for entropy calc
# (HomologMin is needed for entropy calc, so an error won’t occur when dividing by 0)
$ScoreMin       = 100;   # Min value of match score
$BitCutOff      = 40;
$p_directory = "nblast_all";
opendir (DIRECTORY, $p_directory);
while (defined($p_filename = readdir(DIRECTORY))) {
  if ($p_filename != "." || $p_filename != "..") {
      # Initialize all variables
      $PDBName        = "";         # Passed param; used to tag residues in output only
      $Line           = "";
      @LineSplit      = ();
      $ScoreBits      = 0;
      $Expect         = 0;
      $IdenPercent    = 0;
      $PositivePercent= 0;
      $QuerySeq       = "";         # Original Query Seq (from 1st pair)
      $QuerySeqTemp   = "";         # Query Seq in pair for manipulation
      $SubjectSeqTemp = "";         # Subject Seq in pair for manipulation
      $QueryOffset    = 0;          # Initial seq offset (always 1 for 1st seq)
      $LengthDiff     = 0;          # Temp vars ...
      $QueryOffDummy  = 0;
      $Dummy          = "";
      $Dummy2         = "";
      @SeqList        = ();
      $LineOut        = "";
      %EntropyCount   = ();
      $EntropyCountTot= 0;
      $Prob           = 0;
      $Entropy        = 0;
      # Setup files for writing and reading
   open ( IN  ,"$p_directory/$p_filename") or die "Cannot open input files for read"."\n";
      $PDBName = substr($p_filename,0,4);         # Get PDB protein name as 1st parameter
     open ( OUT ,’>’.$PDBName."_".$BitCutOff.".ent") or die "Cannot open output file for write.\n";
   open ( OUTD ,’>’.$PDBName."_".$BitCutOff.".dbg") or die "Cannot open debug file for write.\n";
      while( <IN> ) {
        $Line = $_ ; chomp($Line);                 # Save current line
        # ----------Extracting BitScore, Expected Value, Identity, Positives----------
        while ( $Line =˜ /ˆ Score =/ ) {           # Find next set of Query/Sbjct
            $QueryOffset    = 0 ;                   # Reset values
            $QuerySeqTemp   = "";
            $SubjectSeqTemp = "";
            @LineSplit = split(/ +/, $Line);
```

```
$ScoreBits   = $LineSplit[3];
$Expect      = $LineSplit[8];           # Extract Expect value
if ($Expect =˜ /ˆe/) { $Expect = "1".$Expect }; # If format "e-xxx" add ’1’ prefix
$Line = <IN>; chomp($Line);             # Save next line
@LineSplit = split(/ +/, $Line);
$IdenPercent = $LineSplit[4];           # Extract Identity%, stripping
$IdenPercent =˜ s/[(),%]//g;            # strip out of "(xxx%)," format
$IdenPercent = $IdenPercent / 100.0;
$PositPercent= $LineSplit[8];           # Same with Positive%
$PositPercent =˜ s/[(),%]//g;
$PositPercent = $PositPercent / 100.0;
$Line = <IN>;                           # Skip next line (should be blank)
$Line = <IN>;
# ----------Extracting Query Sequences, Subject Sequences, Query Offsets----------
while ( $Line =˜ /ˆQuery  / ) {         # Find 1st Query line in set
  ($Dummy, $QueryOffDummy, $Line, $Dummy2) = split(/ +/, $Line);    # Separate into fields
  if ($QueryOffset == 0) {             # Keep 1st $QueryOffset
    $QueryOffset = $QueryOffDummy };
  $QuerySeqTemp = $QuerySeqTemp.$Line;  # Combine running seq
  $Line = <IN>;                         # Throw away 2nd line
  $Line = <IN>; chomp($Line);           # Save 3rd = Sbjct line
  $Line =˜ s/ˆSbjct  \d+ +//;           # Strip seq prefix
  $Line =˜ s/ *\d+.*$//;                 # Strip suffix
  $SubjectSeqTemp = $SubjectSeqTemp.$Line; # Combine running seq
  $Line = <IN>;                         # Throw away 2nd line
  $Line = <IN>; chomp($Line);           # Next line (another Query?)
};
if ( $QuerySeq eq "") {                 # Very 1st Query is saved
  $QuerySeq = $QuerySeqTemp;
  $QuerySeq =˜ s/X/x/g;                 # However, convert special X,N chars to lower case
  # $QuerySeq =˜ s/N/n/g;               # (Should be removed, N is used for nucleotides only)
  $QuerySeq =˜ s/-//g;                  # Also, extract insertations (’-’)
  $ScoreMin = $ScoreBits*$BitCutOff/100;
  printf OUTD "# === Original Sequence (from 1st match)\n";
  printf OUTD $QuerySeq . "\n\n";
  printf OUTD "MaxScore = ";
  printf OUTD $ScoreBits;
  printf OUTD ", MinScore = ";
  printf OUTD $ScoreMin . "\n\n";
  push(@SeqList, $QuerySeq."\n");      # Storing Sequences for Entropy calculations
};
# =====  WRITE DEBUG FILE  =====
printf OUTD "# === NEW MATCH PAIR   Offset = " . $QueryOffset . "\n";
```

```
      printf OUTD $QuerySeqTemp   . "\n";     # Write out complete seq
      printf OUTD $SubjectSeqTemp . "\n\n";
      # =====  PROCESS SEQUENCES =====
      # Fill out prefix offset
      $QuerySeqTemp = substr($QuerySeq, 0, -1+$QueryOffset) . $QuerySeqTemp;
      $SubjectSeqTemp = ("-" x (-1+$QueryOffset)) . $SubjectSeqTemp;
      printf OUTD "# === Matched pair with prefix & suffix filled\n";
      printf OUTD $QuerySeqTemp   . "\n";     # Write out complete seq
      printf OUTD $SubjectSeqTemp . "\n\n";
      # Find & delete insertions
      for ($i = -1+length($QuerySeqTemp); $i >= 0; $i += -1) {
        if ( substr($QuerySeqTemp, $i, 1) eq "-" ) {
          substr($QuerySeqTemp, $i, 1) = "";
          substr($SubjectSeqTemp, $i, 1) = "";};};
      # Fill out suffix if necessary
      $LengthDiff = length($QuerySeq) - length($SubjectSeqTemp);
      if ( $LengthDiff > 0 ) {
        $QuerySeqTemp   = $QuerySeqTemp . substr($QuerySeq, - $Length);
        $SubjectSeqTemp = $SubjectSeqTemp . ("-" x $LengthDiff);};
      # =====  QUALIFY THIS SEQUENCE  =====
      # if ($IdenPercent < $IdenPercentMin) {   # If not-qualified, flag as lower case
      if ($ScoreBits < $ScoreMin) {   # If not-qualified, flag as lower case
        $SubjectSeqTemp = lc($SubjectSeqTemp);};
      # =====  WRITE OUT SEQUENCE  =====
      printf OUTD "# === Matched pair with insertions omitted\n";
      printf OUTD $QuerySeqTemp   . "\n";     # Write out complete seq
      printf OUTD $SubjectSeqTemp . "\n\n";
      # ===== REMOVE UNQUALIFIED SEQS FROM THE FINAL ENTROPY FILE OUTPUT =====
      if ($ScoreBits >= $ScoreMin) {
        push(@SeqList, $SubjectSeqTemp."\n");};}; };
close(IN);
for ($i = 0; $i <= -2+length( @SeqList[0] ); $i += 1) { # Why is the -2 value there?
   $LineOut = "";
   for ($j = 1; $j <= $#SeqList; $j += 1) {
      $LineOut = $LineOut . substr( @SeqList[$j], $i, 1); # Cycling thru seq j, at pos i
         };
   # =====  COMPUTE ENTROPY  =====
   $Line = $LineOut;
   $Line =˜ s/[ˆA-Z]//g;                # Delete anything not capital (IMPORTANT!!!!)
   @LineSplit = split("", $Line);
   %EntropyCount    = ();
   $EntropyCountTot = 0;
   foreach $i (@LineSplit) {
    if(($i eq "A") || ($i eq "V") || ($i eq "L") || ($i eq "I") || ($i eq "M") ||($i eq "C")){
            $category = "aliphatic";
            $EntropyCount{$category}++ ;
      #print "i is $i and count in aliphatic is $EntropyCount{$category} \n";}
    elsif( ($i eq "F") || ($i eq "W") || ($i eq "Y") ||($i eq "H")){
   $category = "aromatic";
   $EntropyCount{$category} = $EntropyCount{$category} + 1 ;}
    elsif( ($i eq "S") || ($i eq "T") || ($i eq "N") ||($i eq "Q")){
   $category = "polar";
    $EntropyCount{$category} = $EntropyCount{$category} + 1 ;}
    elsif( ($i eq "K") || ($i eq "R") ){
   $category = "positive";
    $EntropyCount{$category} = $EntropyCount{$category} + 1 ;}
    elsif( ($i eq "D") || ($i eq "E") ){
$category = "negative";
    $EntropyCount{$category} = $EntropyCount{$category} + 1 ;}
            elsif( ($i eq "G") || ($i eq "P") ){
$category = "special";
    $EntropyCount{$category} = $EntropyCount{$category} + 1 ;}
            $EntropyCountTot++ ;}
         @AllEntValues = sort(values(%EntropyCount));
         if ($EntropyCountTot >= $HomologMin) {
            $Entropy = 0;
            for ($j = 0; $j <= $#AllEntValues; $j++) {
              # print "entropy value is $AllEntValues[$j] \n";
             #  print "Total count is $EntropyCountTot \n";
               $Prob = $AllEntValues[$j] / $EntropyCountTot ;
               $Entropy = $Entropy - ($Prob * ( log($Prob)/log(2) ));}
            # Debug Code
            printf OUTD "@"
               ." Entropy= "        .$Entropy
               ." EntropyCount= "   .join(" ",%EntropyCount)
               ." EntropyCountTot= ".$EntropyCountTot
               ." AllEntValues= "   .join(" ",@AllEntValues)
               ."\n";}
         else {
            $Entropy = -1;                    # Flag as error -- too few homologs
         };
         $LineOut = "E= ".sprintf("% .3f",$Entropy)
                  ." A= ".$LineOut;
         # =====  CREATE OUTPUT LINE PREFIX  =====
         $LineHeader = "D "                         # Start line format eg "D 1agm_001_A"
                       .$PDBName."_"
```

```
                        .sprintf("%03d",1+$i)
                        ."_". substr(@SeqList[0], $i, 1);
         printf OUT $LineHeader ." ". $LineOut . "\n";};
      close(OUTD);
      close(OUT);};};
```


```
----------------------------------------------------------------------------------------------------
#Author: Radhika Pallavi Mishra
#Date: June 6, 2008
#Purpose: Download '.HSSP' files from
#"ftp.embl-heidelberg.de/pub/databases/protein_extras/hssp" according to a protein list.
#File: ftp-scriptHSSP.pl
------------------------------------------------------------------------------------------------------
#!/usr/bin/perl -w
use Net::FTP;
open(FHNDL,"SeqIDlearningset268.txt");
$line = <FHNDL>;
print "filename is $line\n";
$ftp = Net::FTP->new("ftp.embl-heidelberg.de");
$ftp->login("anonymous",'-anonymous@');
$ftp->cwd("/pub/databases/protein_extras/hssp");
while(!eof(FHNDL)){
$line = lc($line);
chomp($line);
$filetoftp = $line."\.hssp";
$ftp->get($filetoftp,$filetoftp);
$line = <FHNDL>;
print "$line\n";
}
$ftp->quit;
close(FHNDL);
```


```
------------------------------------------------------------------------------------------------------
# Author: Radhika Pallavi Mishra
# Date : September 21, 2008
# Purpose: Program to extract entropy values from entropy "ENT Files" and compute
#fractional analysis print in one file for aggregate plot
# File: extract_fractanalysis_entropy_aggr.pl
------------------------------------------------------------------------------------------------------
if($#ARGV < 0) {
```

```
print "Usage : perl extract_fractanalysis_entropy_aggr.pl <directory with .ent files>\n";
        exit;}
# Open a directory and read a file
$p_directory = $ARGV[0] ;
opendir (DIRECTORY, $p_directory) or die "cannot open";
while (defined($p_filename = readdir(DIRECTORY))) {
if ($p_filename =˜ /\.ent/) {
  $filetoopen = $p_directory."/".$p_filename;
  print "file to open is $filetoopen\n";
  open(FHNDL, $filetoopen) or die "Cannot open file $p_filename";
  $output_filename = substr($p_filename,0,5)."\.fract";
  open(OUTFHNDL, ">$output_filename");
  $i=0;
  #print "$lines";
  do {
   $lines = <FHNDL>;
    @filearray = split(/ +/,$lines); # @LIST = split(/PATTERN/, STRING);
$entropy_val = $filearray[3];
    chomp($entropy_val);
    #split the alignments
      $alignment= $filearray[5];
      chomp($alignment);
       @align= split(//,$alignment);
   $totalLength = $#align + 1;
       # calculate gapfraction -
       $numgaps = 0;
   $gapfraction = 0;
   # calculate fraction small residues (Alanines A and Glycines G)
       $small_residues = 0;
       $small_residues_fraction = 0;
       # calculate fraction strongly hydrophobic (V, I, L, F, Y, M, W)
       $strongly_hydrophobic = 0;
   $strongly_hydrophobic_fraction = 0;
# calculate fraction strongly hydrophobic with gaps= fraction str. hydrophobic- fractiongaps
#fraction of small residues with gaps =
# Sequence entropy with gaps = average sequence entropy- fraction gaps
     foreach $amino_acid (@align) {
          if($amino_acid eq "-"){
  $numgaps = $numgaps +1 ;}
if(($amino_acid eq "A") || ($amino_acid eq "G")){
$small_residues = $small_residues +1 ;}
  if(($amino_acid eq "V") || ($amino_acid eq "I")|| ($amino_acid eq "L")
||($amino_acid eq "F")|| ($amino_acid eq "Y") || ($amino_acid eq "M")|| ($amino_acid eq "W") ){
```

```
$strongly_hydrophobic = $strongly_hydrophobic +1 ;}}
     $num_non_gap_amino_acids = $totalLength - $numgaps;
        if ($num_non_gap_amino_acids > 0){
    $gapfraction = $numgaps/$num_non_gap_amino_acids ;
        $small_residues_fraction = $small_residues/$num_non_gap_amino_acids;
    $strongly_hydrophobic_fraction = $strongly_hydrophobic/$num_non_gap_amino_acids;
        $non_strongly_hydrophobic_fraction = 1- $strongly_hydrophobic_fraction;}
        else{
          $gapfraction = $num_non_gap_amino_acids;
            $small_residues_fraction = $num_non_gap_amino_acids;
    $strongly_hydrophobic_fraction = $num_non_gap_amino_acids;
            $non_strongly_hydrophobic_fraction = $num_non_gap_amino_acids;}
print OUTFHNDL"E=$entropy_val,FG=$gapfraction,FSR=$small_residues_fraction,";
print OUTFHNDL"FSHP=$strongly_hydrophobic_fraction,FNSHP=$non_strongly_hydrophobic_fraction\n";
   } while(!eof(FHNDL));
   print OUTFHNDL "\n\n";
   close(FHNDL);
   close(OUTFHNDL);}}
end;

----------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------

----------------------------------------------------------------------------------------------------------
# Author: Radhika Pallavi Mishra
# Date : September 21, 2008
# Purpose: Program to extract entropy values from entropy "ENT Files" and
#print in one file for aggregate plot
# File: extract_individualfractentropy_density_aggr.pl
----------------------------------------------------------------------------------------------------
if($#ARGV < 1) {
print "Usage : perl extract_entropy_aggr.pl <directory with .fract files>
<directory with density files> \n";
}
# Open a directory and read a file
$p_directory = $ARGV[0] ;
$p1_directory = $ARGV[1] ;
# Open files for reading and writing
  # open file from density directory
  opendir (DIRECTORY1, $p1_directory) or die "cannot open directory $p1_directory \n";
    while (defined($p1_filename = readdir(DIRECTORY1))){
     if ($p1_filename != "." || $p1_filename != "..") {
       print "filename is $p1_filename and directory is $p1_directory \n";
  open ( FHNDL1  ,"$p1_directory/$p1_filename") or die "Cannot open input files for read"."\n";
       $PDBName = substr($p1_filename,0,5);         # Get PDB protein name as 1st parameter
```

```
        $PDBName = uc($PDBName);
  open(OUTFHNDL, ">$PDBName\.txt");
  # open  corresponding fract file
    $filetoopen = $p_directory."/".$PDBName."\.fract";
    print "file to open is $filetoopen\n";
    open(FHNDL, $filetoopen) or die "Cannot open file $p_filename";
    $lines1 = <FHNDL1>;
    $lines = <FHNDL>;
    do{
   if ($lines1 =˜ /C\(9\)/){
        @filearray1 = split(/ +/,$lines1); # @LIST = split(/PATTERN/, STRING);
     $density_val = $filearray1[9];
          $entropy_val = $lines;
          chomp($density_val);
          chomp($entropy_val);
           print OUTFHNDL "Den=$density_val,$entropy_val\n";
    $lines= <FHNDL>;}
$lines1 = <FHNDL1>;}
  while(!eof(FHNDL1) && !eof(FHNDL));
 print OUTFHNDL "\n\n";
close(FNDHL1);
close(FHNDL);
close (OUTFHNDL);
}}
end;

---------------------------------------------------------------------------------------------
---------------------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------
#Author : Radhika Pallavi Mishra
#Date : September 21, 2008
#Purpose : Program to calculate aggregate values of all the parameters for each protein.
#Takes input from the .text files and prints output in separate directory.
#File: calculate_aggr_per_protein.pl
-------------------------------------------------------------------------------------------

if($#ARGV < 1) {
print "Usage : perl calculate_aggr_per_protein.pl <Input directory with .txt files with
individual protein density-parameter value> <output directory name> \n";
       exit;
}
# Open a directory and read a file
$p_directory = $ARGV[0] ;
$p_directory1 = $ARGV[1];
```

```
# Open files for reading and writing
   # open file from density directory
  opendir (DIRECTORY, $p_directory) or die "cannot open directory $p1_directory \n";
    while (defined($p_filename = readdir(DIRECTORY))){
     if ($p_filename =˜ /txt/) {
       print "filename is $p1_filename and directory is $p1_directory \n";
       open ( FHNDL  ,"$p_directory/$p_filename") or die "Cannot open input files for read"."\n";
       $PDBName = substr($p_filename,0,4);         # Get PDB protein name as 1st parameter
       $outputfilename = $PDBName."Aggr"."\.txt";
 $outputfilename = $p_directory1."/".$outputfilename ;
     open(OUTFHNDL, ">$outputfilename");
      # Following Arrays will store the average value of ent, fg, fsr fshp and fnshp
      #for den = 0 to 40 in their index 0 to 40
     my @aggr_rsa_array = ();
  # Following Array will store the average number of occurences of den=i, at index i
         my @num_density_occurences = ();
       for($i=0;$i<=40;$i++){
  $aggr_rsa_array[$i] = 0;
  $num_density_occurences[$i] = 0;}
  do{
$lines = <FHNDL>;
if($lines =˜ /ˆPDB/){
 @filearray = split(/,/,$lines); # @LIST = split(/PATTERN/, STRING);
  print "$lines ";
  @temp = split(/\=/,$filearray[2]);
      $den = $temp[1];
      print "Density is $den \n";
      if($den ne "NA"){
  for($i=0;$i<=40;$i++) {
if($den == $i){
@temp = split(/=/,$filearray[3]);
$rsa = $temp[1];
print "$rsa\n";
                $aggr_rsa_array[$i] = $aggr_rsa_array[$i] + $rsa;
  $num_density_occurences[$i] =  $num_density_occurences[$i] + 1;}}}}}
  while(!eof(FHNDL));
for($i=0; $i<=40; $i++){
 if($num_density_occurences[$i] > 0) {
 $aggr_rsa = $aggr_rsa_array[$i] / $num_density_occurences[$i] ;
 print OUTFHNDL "Den=$i,RSA=$aggr_rsa,N=$num_density_occurences[$i]\n";}
 else{
 print OUTFHNDL "Den=$i,RSA=NA,N=0\n";}}
 print OUTFHNDL "\n\n";
close(FHNDL);

close (OUTFHNDL);}}

end;

----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------

----------------------------------------------------------------------------------------------------
# Author: Radhika Pallavi Mishra

# Date: September 21, 2008

# Purpose: Program to calculate double aggregate values of parameters

#from the individual protein aggregate files('.txt').

# File: double_aggr_forPlot.pl
--------------------------------------------------------------------------------------------------
if($#ARGV < 1) {

print "Usage : perl extract_entropy_aggr.pl <directory with .txt files> <name of output file> \n";}

# Open a directory and read a file

$p_directory = $ARGV[0] ;

$outputfile = $ARGV[1];

# Open files for reading and writing

  open(OUTFHNDL, ">$outputfile");

  # open file from density directory

  opendir (DIRECTORY, $p_directory) or die "cannot open directory $p_directory \n";

    while (defined($p_filename = readdir(DIRECTORY))){

     if ($p_filename != "." || $p_filename != "..") {

       print "filename is $p_filename and directory is $p_directory \n";

       open ( FHNDL  ,"$p_directory/$p_filename") or die "Cannot open input files for read"."\n";

  $lines = <FHNDL>;

  #print "$lines";

  do

  {

       print OUTFHNDL "$lines";

#print "$lines\n";

        $lines= <FHNDL>;}

while(!eof(FHNDL));}}

close(FHNDL);

close(OUTFHNDL);

end;

-------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------
Author: Radhika Pallavi Mishra

Date: November 6, 2008

Purpose:Program to list the number of alignments in a blast file saved as .txt
```

```
File: listNoAlignments.pl
-------------------------------------------------------------------------------------------------
if($#ARGV < 2) {
print "Usage : perl listNoAlignments.pl <directory with .txt files> <name of output file> \n";
}
# Open a directory and read a file
$p_directory = $ARGV[0] ;
$output = $ARGV[1];
open(OUTFHNDL, ">$output");
opendir (DIRECTORY, $p_directory) or die "cannot open";
while (defined($p_filename = readdir(DIRECTORY))) {
if ($p_filename =˜ /\.txt/) {
  $filetoopen = $p_directory."/".$p_filename;
  #print "file to open is $filetoopen\n";
  open(IN, $filetoopen) or die "Cannot open file $p_filename";
  #print" opened file $p_filename from $p_directory\n";
  $line = <IN>;
  $n=0;
  while(!eof(IN))
    {
      if ( $line =˜ /ˆ Score =/) {
          $n++;
          }
$line = <IN>;}
print "$n\n";
print OUTFHNDL "Total Number of alignments = $n\n";
close(IN);}
close(OUTFHNDL);
end;

--------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------------
#Author:Radhika Pallavi Mishra
#Date: September 15, 2008
#Purpose: Program to list the number of residues in a density file
#File: No_of_res_count.pl
---------------------------------------------------------------------------------------------------
if($#ARGV < 2) {
print "Usage : perl extract_entropy_aggr.pl <directory with .den files> <name of output file> \n";
}
# Open a directory and read a file
$p_directory = $ARGV[0] ;
$outputfile = $ARGV[1];
```

```
# Open files for reading and writing
  open(OUTFHNDL, ">$outputfile");
  # open file from density directory
  opendir (DIRECTORY1, $p_directory) or die "cannot open directory $p_directory \n";
    while (defined($p_filename = readdir(DIRECTORY1))){
      if ($p_filename != "." || $p_filename != "..") {
        print "filename is $p_filename and directory is $p_directory \n";
    $filetoopen = $p_directory."/".$p_filename;
  print "file to open is $filetoopen\n";
     open(FHNDL, $filetoopen) or die "Cannot open file $p_filename";
$lines = <FHNDL>;
  while(!eof(FHNDL)){
        if ($lines =˜ /NO_RESIDUES=/){
print "$lines \n";
        @filearray = split(/ +/,$lines);
$no_res = $filearray[2];
        chomp($no_res);
        print OUTFHNDL "$p_filename $no_res\n";}
$lines= <FHNDL>;}
close(FHNDL);}}
 close(OUTFHNDL);
end;

-------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------
# Author: Radhika Pallavi Mishra
# Date : November 11, 2008
# Purpose: Program to list the scores in a blast file saved as .txt, and the corresponding bit score
# File: Bitscorelistno_ofsubject.pl
-----------------------------------------------------------------------------------------------------

if($#ARGV < 2) {
print "Usage : perl Bitscorelistno_ofsubject.pl <directory with .txt files> <name of output file> \n";}
# Open a directory and read a file
$p_directory = $ARGV[0] ;
$output = $ARGV[1];
open(OUTFHNDL, ">$output");
opendir (DIRECTORY, $p_directory) or die "cannot open";
while (defined($p_filename = readdir(DIRECTORY))) {
if ($p_filename =˜ /\.txt/) {
  $filetoopen = $p_directory."/".$p_filename;
  open(IN, $filetoopen) or die "Cannot open file $p_filename";
  $line = <IN>;
  while(!eof(IN)){

      if ( $line =˜ /ˆ Score =/) {

          print "$line \n";

          @filearray = split(/ +/,$line);

          $bitscore = $filearray[3];

          chomp $bitscore;

          print "$bitscore\n";

           print OUTFHNDL "$bitscore\n";}

$line = <IN>;}

close(IN);}}

close(OUTFHNDL);

end;

----------------------------------------------------------
----------------------------------------------------------
----------------------------------------------------------
```

Appendix B

Additional Figures

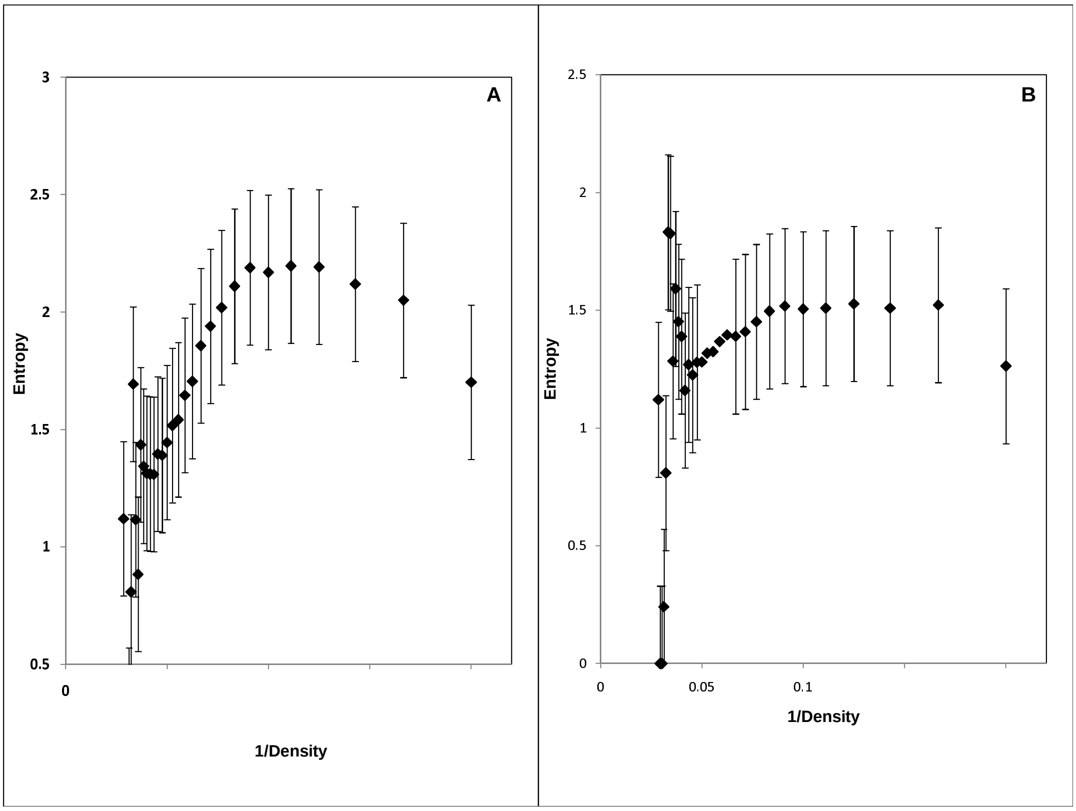


Figure B.1: Aggregate entropy-inverse density correlation plots of 268 proteins for a set of BLASTP alignment cutoffs. Here, double average entropy is calculated by averaging the entropy values per protein and then averaging them over the 268 set of proteins. Entropy-inverse density plot for A. Expect cutoff 0.001. B. Percentage Identity 25%. In both these plots, gaps are excluded in the entropy calculations and entropy values are averaged for each inverse density value.

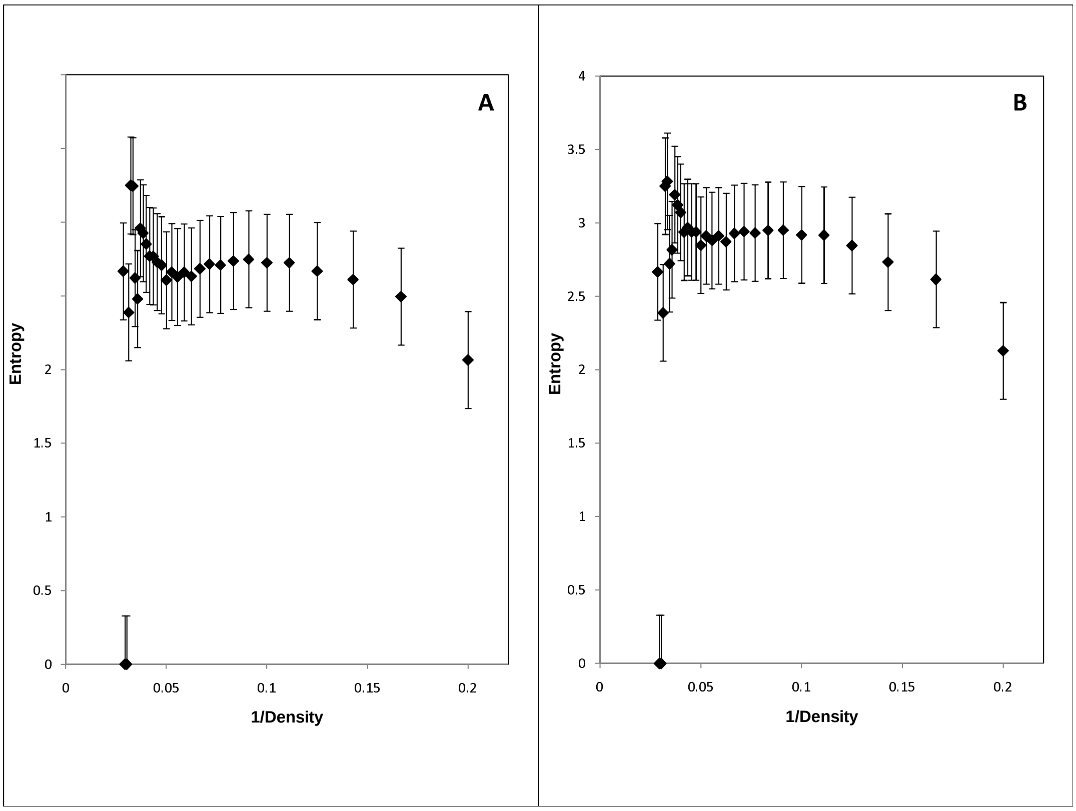
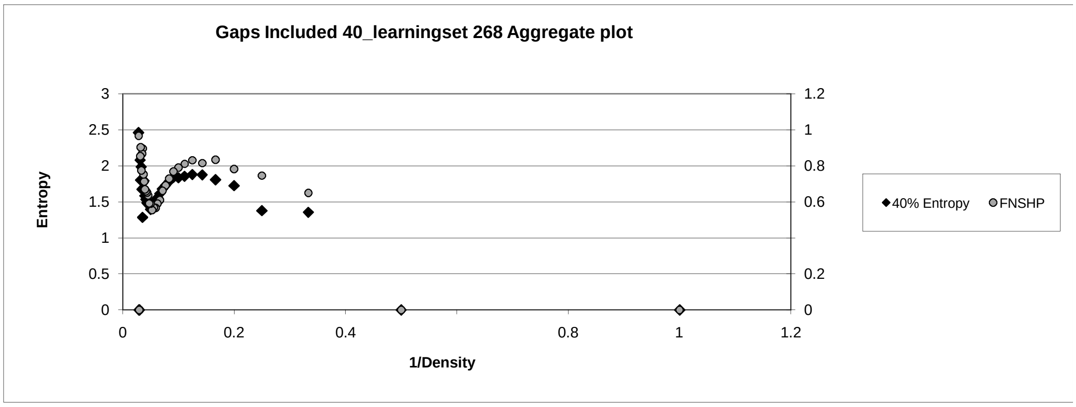


Figure B.2: Entropy-inverse density correlation plots. A. For Expect 0.001 B. For Percentage Identity 25%. In both these cases, mutational insertions and deletions, represented by gaps, were included in the entropy calculations as the 21st amino acid.



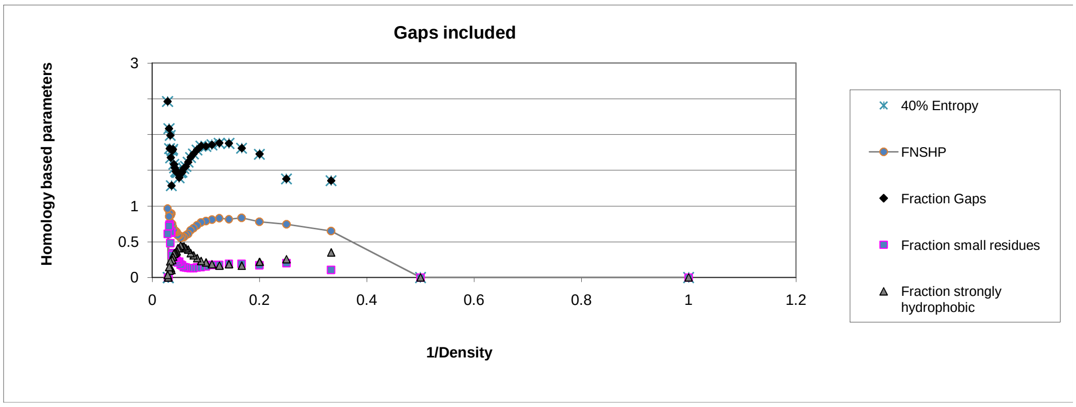
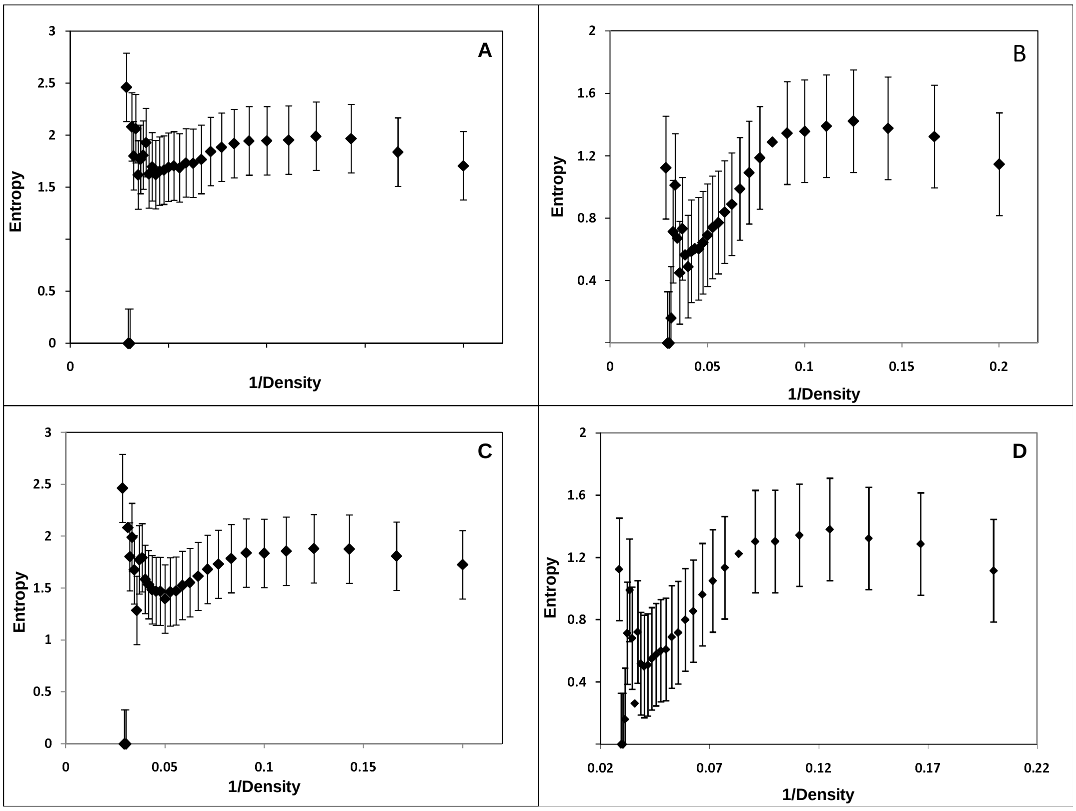
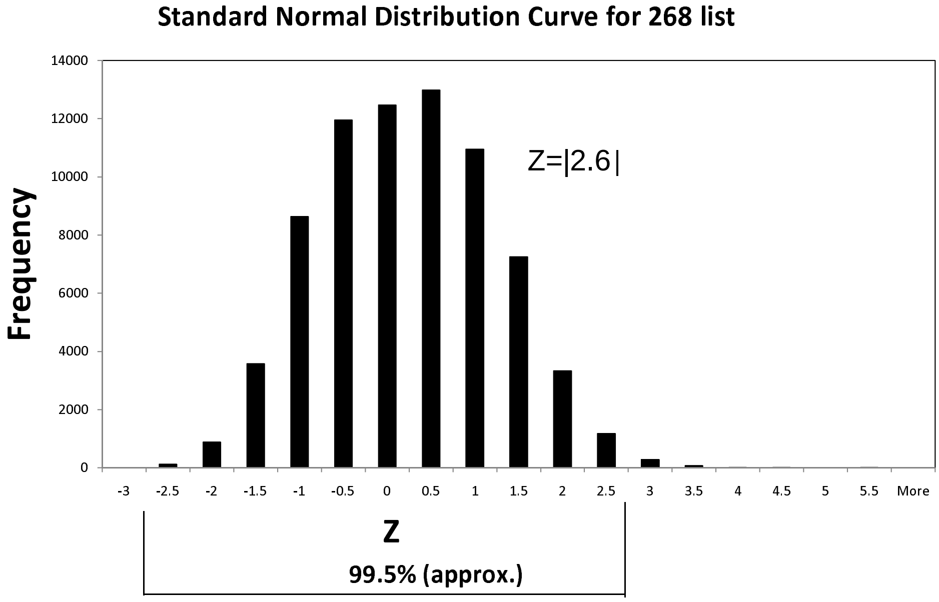


Figure B.3: Gaps Included Aggregate plots for learning set list.

Figure B.4: Entropy inverse density plots for a set of 268 proteins. A. Single average aggregate plot of 40% entropy calculated by considering gap as the $21st$ term. B. Single average aggregate 40% entropy- inverse density plot calculated by excluding gaps in the entropy calculations. C. Double average aggregate 40% entropy- inverse density plot calculated for each packing density position. Here, gaps were taken into account while calculating the Entropy values. D. Double average aggregate plot of gaps excluded 40% entropy for a set of 75891 query proteins.





Packing Density : 1-3 and 26-35 Anomalous Regions
Packing Density: 4-10 Major Region II
Packing Density: 11-25 Major Region I

Figure B.5: Z Score analysis of the learning set list of proteins.

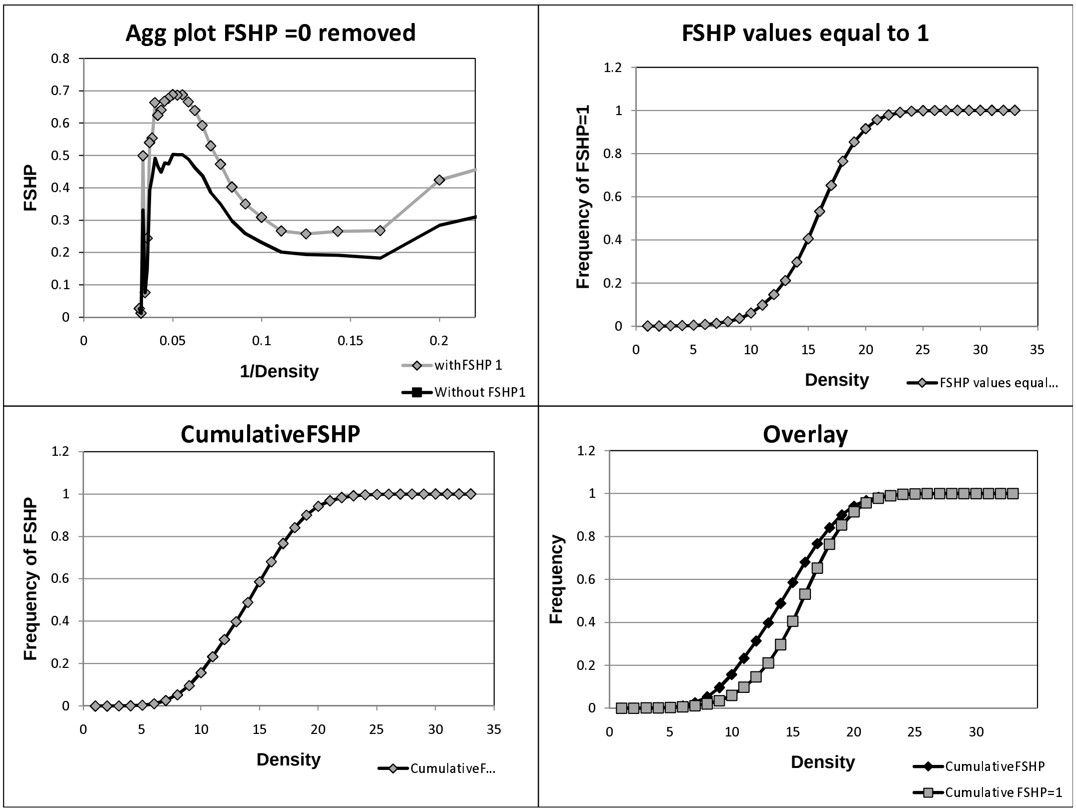


Figure B.6: Aggregate correlation plot of fraction strongly hydrophobic for the learning set list. Aggregate correlation plot of fraction strongly hydrophobic for the learning set list versus inverse density, for residues that have exclusively strongly hydrophobic residues in their alignments (FSHP = 1) and residues that have some other residues in addition to strongly hydrophobic residues in their alignments (Without FSHP 1). All this data was calculated after removing all the residues with FSHP = 0 .

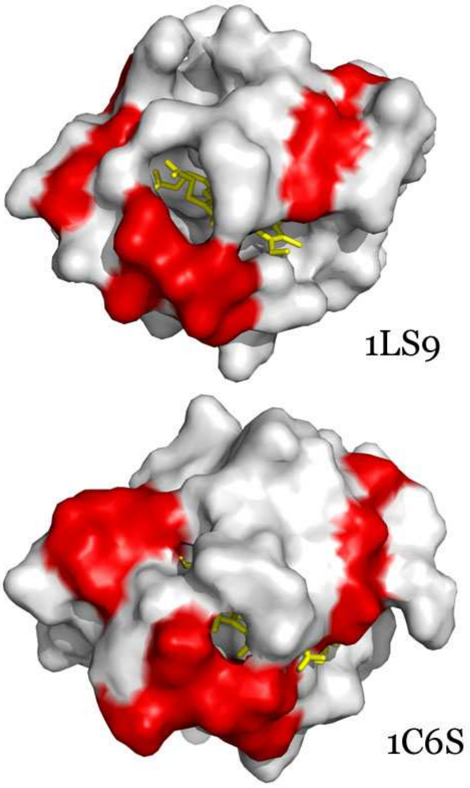
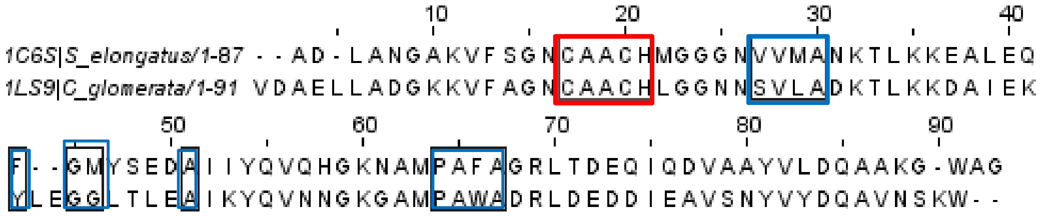


Figure B.7: Visual alignment of 1LS9 and 1C6S. Both Cytochrome C6. Shown are the positions (red) of the literature-noted patch residues from 1C6S and aligned for 1LS9 (Beissenger et al., 2004; Do, S., San Jose State University, personal communication, 2009). Note 1LS9 is an example of test protein for application of filter parameters. Examination of prediction of surface hydrophobic patches should prove interesting.



```
                                           10        20        30        40
1C6S|S_elongatus/1-87   - - A D - L A N G A K V F S G N C A A C H M G G G N V V M A N K T L K K E A L E Q
1LS9|C_glomerata/1-91   V D A E L L A D G K K V F A G N C A A C H L G G N N S V L A D K T L K K D A I E K

         50        60        70        80        90
F - - G M Y S E D A I I Y Q V Q H G K N A M P A F A G R L T D E Q I Q D V A A Y V L D Q A A K G - W A G
Y L E G G L T L E A I K Y Q V N N G K G A M P A W A D R L D E D D I E A V S N Y V V Y D Q A V N S K W - -
```

Figure B.8: Alignment of 1C6S and 1LS9, two related proteins by ClustalW in Jalview 2.4. The conserved motif for Cytochrome C6 is bordered in red and the literature-noted patch residues for 1C6S are bordered in blue (Do, S., San Jose State University, personal communication, 2009).

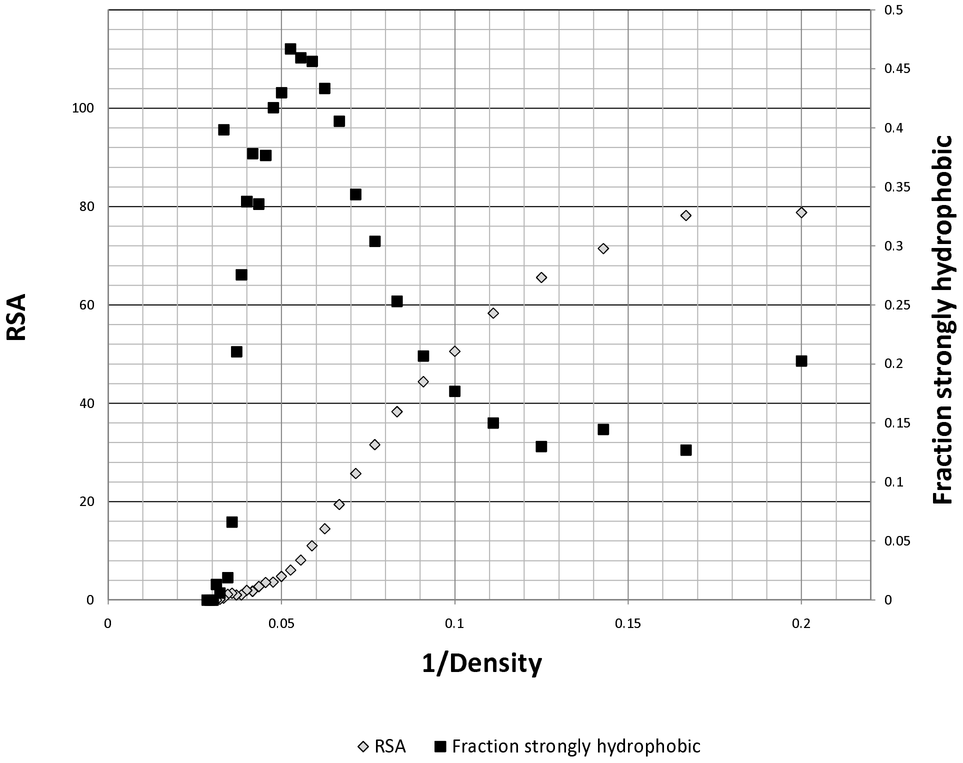


Figure B.9: RSA versus Fraction strongly hydrophobic for the learning set list of 268 proteins.

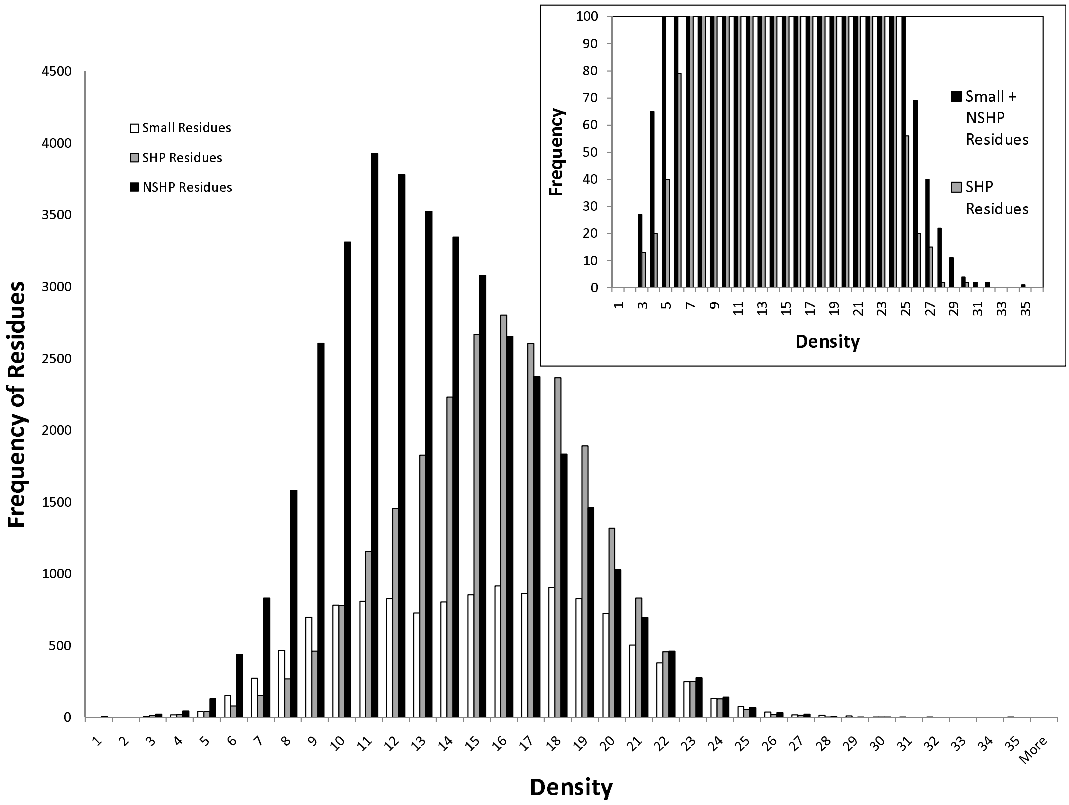


Figure B.10: Frequency distribution of query residues for the learning set list of 268 proteins. Enlarged image of the frequency distribution is shown in the insert.

Appendix C

Additional Tables

Table C.1: List of 65 monomeric protein chains present in the learning set list.

IDs = PDB ID + Chain	IDs = PDB ID + Chain	IDs = PDB ID + Chain
13PKA	1G2AA	1THTA
1AG9A	1GARA	1TOAA
1AH7A	1GJMA	1URPA
1AKOA	1HF8A	1VBTA
1AMUA	1ILR1	1XGSA
1ATLA	1KPTA	256BA
1AW7A	1KWAA	256LA
1AYLA	1MPGA	2ACYA
1BEAA	1NAWA	2ATJA
1BINA	1NP4A	2BC2A
1BYOA	1PBGA	2BLSA
1C02A	1PDAA	2G3PA
1CKIA	1QAZA	2IHLA
1CQXA	1QCIA	2MBRA
1DYSA	1QHAA	2SCPA
1EHYA	1QJPA	2SHPA
1EWFA	1QMEA	2TPSA
1FEHA	1QPAA	2UGIA
1FGKA	1QTQA	3PMGA
1FJMA	1RHSA	830CA
1FKDA	1RNEA	8PTIA
1FMTA	1SHKA	

Table C.2: List of 99 homodimeric protein chains present in the learning set list.

IDs = PDB ID + Chain	IDs = PDB ID + Chain	IDs = PDB ID + Chain	IDs = PDB ID + Chain	IDs = PDB ID + Chain	IDs = PDB ID + Chain
12ASA	1BBHA	1CSHA	1ISAA	1QR2A	2ILKA
1A4IA	1BD0A	1CTTA	1IVYA	1REGX	2NACA
1A4UA	1BIQA	1CZJA	1JHGA	1RPOA	2OHXA
1AA7A	1BISA	1DAAA	1JSGA	1SESA	2SPCA
1ADEA	1BJWA	1DORA	1KBAA	1SLTA	2SQCA
1AFWA	1BMDA	1DPGA	1KPFA	1SMNA	2TCTA
1AJSA	1BRWA	1DQSA	1M6PA	1SMTA	2TGIA
1AMKA	1BSLA	1E98A	1MKBA	1SOXA	3DAPA
1AORA	1BUOA	1EBHA	1MORA	1TC1A	3GRSA
1AQ6A	1BXGA	1F13A	1NOXA	1TOXA	3SDHA
1AUOA	1BXKA	1FIPA	1NSEA	1TRKA	5CSMA
1B3AA	1CDCA	1FROA	1NSYA	1UBYA	5RUBA
1B5EA	1CG2A	1GVPA	1OACA	1UTGA	8PRKA
1B67A	1CHMA	1HJRA	1OPYA	1VOKA	9WGAA
1B8AA	1CMBA	1HXPA	1PGTA	1XSOA	
1B8JA	1CNZA	1ICWA	1QFHA	2ARCA	
1BAMA	1COZA	1IMBA	1QHIA	2HDHA	

Table C.3: List of 40 heterodimeric protein chains present in the learning set list.

IDs = PDB IDs +Chain	IDs = PDB IDs +Chain	IDs = PDB IDs +Chain	IDs = PDB IDs +Chain
1A2KA	1DFJI	1GOTG	1TX4B
1AK4C	1DHKA	1GUAB	1YCSA
1AVWB	1DHKB	1HIAI	1YCSB
1BRSA	1EFNB	1HWGA	1YDRE
1BRSD	1EFUA	1HWGB	2PCCA
1CSEE	1EFUB	1MCTA	2SICI
1CSEI	1FINB	1NMBN	2TRCP
1DANL	1FLEI	1OSPO	3SGBE
1DANT	1GOTA	1STFI	3SGBI
1DANU	1GOTB	1TX4A	4HTCI

Table C.4: Z Score values at each density position for the learning set list of proteins.

Packing Density	Z=(x-mean)/std.dev
3	-2.850413391
4	-2.600950285
5	-2.351487178
6	-2.102024072
7	-1.852560965
8	-1.603097858
9	-1.353634752
10	-1.104171645
11	-0.854708539
12	-0.605245432
13	-0.355782326
14	-0.106319219
15	0.143143887
16	0.392606994
17	0.642070101
18	0.891533207
19	1.140996314
20	1.39045942
21	1.639922527
22	1.889385633
23	2.13884874
24	2.388311847
25	2.637774953
26	2.88723806
27	3.136701166
28	3.386164273
29	3.635627379
30	3.885090486
31	4.134553593
32	4.384016699
33	4.633479806
34	4.882942912
35	5.132406019

Table C.5: Trend comparison for the three protein lists with the aggregate trend of the learning set list of 268 proteins.

	Entropy		FSHP		FNSHP		FSR		FG	
	Average Ratio	Absolute Deviation	Average Ratio	Absolute Deviation	Average Ratio	Absolute Deviation	Average Ratio	Absolute Deviation	Average Ratio	Absolute Deviation
268/Monomeric	0.996	0.004	1.065	0.065	1.008	0.008	1.053	0.053	1.026	0.026
268/Homodimeric	1.054	0.054	0.963	0.037	0.999	0.001	0.988	0.012	1.909	0.909
268/Heterodimeric	1.211	0.211	1.109	0.109	1.024	0.024	1.463	0.463	4.180	3.180

Table C.6: Distributional analysis of the query residues for the learning set list of 268 proteins.

Packing density	Small Residues (SR)	% SR	Strongly hydrophobic residues (SHP)	% SHP	Non-Strongly Hydrophobic residues(NSHP)	%NSHP	(SR/T)%	(SHP/T)%	(NSHP/T)%	Total No. of Res. at each density T = SHP+NSHP
$x < 4$	5	0.006782	13	0.018	27	0.0366	12.5	32.5	67.5	40
$4 \leq x > 11$	2435	3.302591	1803	2.445	11382	15.437	18.468	13.67463	86.3253697	13185
$11 \leq x \geq 25$	9608	13.03133	22052	29.91	38263	51.896	15.9297	36.56139	63.4386139	60315
$25 < x$	86	0.116642	39	0.053	151	0.2048	45.2632	20.52632	79.4736842	190
1	0	0	0	0	0	0	-	-	-	0
2	0	0	0	0	0	0	-	-	-	0
3	5	0.007	13	0.018	27	0.037	12.5	32.5	67.5	40
4	19	0.026	20	0.027	65	0.088	22.353	23.529	76.471	85
5	43	0.058	40	0.054	173	0.235	20.188	18.779	81.221	213
6	153	0.208	79	0.107	591	0.802	22.836	11.791	88.209	670
7	274	0.372	155	0.210	1107	1.501	21.712	12.282	87.718	1262
8	467	0.633	268	0.363	2048	2.778	20.164	11.572	88.428	2316
9	697	0.945	462	0.627	3303	4.480	18.513	12.271	87.729	3765
10	782	1.061	779	1.057	4095	5.554	16.044	15.983	84.017	4874
11	809	1.097	1157	1.569	4737	6.425	13.726	19.630	80.370	5894
12	828	1.123	1456	1.975	4609	6.251	13.652	24.007	75.993	6065
13	728	0.987	1828	2.479	4253	5.768	11.972	30.061	69.939	6081
14	806	1.093	2231	3.026	4154	5.634	12.623	34.941	65.059	6385
15	855	1.160	2670	3.621	3934	5.336	12.947	40.430	59.570	6604
16	916	1.242	2804	3.803	3570	4.842	14.371	43.991	56.009	6374
17	865	1.173	2604	3.532	3239	4.393	14.804	44.566	55.434	5843
18	907	1.230	2367	3.210	2742	3.719	17.753	46.330	53.670	5109
19	828	1.123	1892	2.566	2287	3.102	19.813	45.274	54.726	4179
20	725	0.983	1318	1.788	1754	2.379	23.600	42.904	57.096	3072
21	505	0.685	831	1.127	1200	1.628	24.865	40.916	59.084	2031
22	380	0.515	456	0.618	843	1.143	29.253	35.104	64.896	1299
23	249	0.338	252	0.342	524	0.711	32.088	32.474	67.526	776
24	133	0.180	130	0.176	276	0.374	32.759	32.020	67.980	406
25	74	0.100	56	0.076	141	0.191	37.563	28.426	71.574	197
26	37	0.050	20	0.027	69	0.094	41.573	22.472	77.528	89
27	18	0.024	15	0.020	40	0.054	32.727	27.273	72.727	55
28	15	0.020	2	0.003	22	0.030	62.500	8.333	91.667	24
29	10	0.014	0	0.000	11	0.015	90.909	0.000	100.000	11
30	1	0.001	2	0.003	4	0.005	16.667	33.333	66.667	6
31	2	0.003	0	0	2	0.003	100	0	100	2
32	2	0.003	0	0	2	0.003	100	0	100	2
33	0	0.000	0	0	0	0	-	-	-	0
34	0	0.000	0	0	0	0	-	-	-	0
35	1	0.001	0	0	1	0.0014	100	0	100	1
More	0	0	0	0	0	0	-	-	-	0